计算机科学
COMPUTER SCIENCE

融合多源图特征的Kcore-GCN反欺诈算法研究

刘炜, 宋友, 卓佩妍, 仵伟强, 廉鑫

引用本文

刘炜, 宋友, 卓佩妍, 仵伟强, 廉鑫. 融合多源图特征的Kcore-GCN反欺诈算法研究[J]. 计算机科学, 2024, 51(6A): 230600040-7.

LIU Wei, SONG You, ZHUO Peiyan, WU Weiqiang, LIAN Xin. Study on Kcore-GCN Anti-fraud Algorithm Fusing Multi-source Graph Features [J]. Computer Science, 2024, 51(6A): 230600040-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600040

融合多源图特征的 Kcore-GCN 反欺诈算法研究

刘 炜[1] 宋 友[1] 卓佩妍[1] 仵伟强[2] 廉 鑫[2]
1 北京航空航天大学软件学院 北京 100191
2 渤海银行 天津 300070
(buaaliuwei@buaa.edu.cn)

摘 要 金融欺诈行为给社会带来了许多负面影响,针对金融欺诈行为,多种人工智能与金融反欺诈算法被提出并应用于实际反欺诈业务场景,取得了不错的成绩。这些反欺诈算法或从用户个体的角度进行欺诈检测,或从节点与网络的拓扑关系的角度进行欺诈检测,或通过学习节点的图嵌入式表示进行欺诈检测,出发角度较为局限,无法进行完备的欺诈检测分析。针对上述问题,设计了一种基于融合多源图特征的 Kcore 图卷积神经网络反欺诈算法,该算法的创新性在于能够高效挖掘网络中节点层级的拓扑关系与全局网络层次的拓扑关系来构建宽领域的特征体系,并通过基于 Kcore 算法的图卷积神经网络完成深层次图结构特征的传播与聚合,最终完成欺诈风险的检测。实验效果表明,该方法相较于相关机器学习算法与图神经网络算法在相关评价指标上均有较大的提升,其中较 LightGBM 算法有 12%的 AUC 值提升,较 GCN 算法有 6%的 AUC 值提升。

关键词:机器学习;图表示学习;图神经网络;金融欺诈检测
中图分类号 TP183

Study on Kcore-GCN Anti-fraud Algorithm Fusing Multi-source Graph Features

LIU Wei[1],SONG You[1],ZHUO Peiyan[1],WU Weiqiang[2] and LIAN Xin[2]
1 College of Software,Beihang University,Beijing 100191,China
2 China Bohai Bank Co.,Ltd,Tianjin 300070,China

Abstract Financial fraud has brought many negative impacts to society,and a variety of AI and financial anti-fraud algorithms have been applied to practical anti-fraud business scenarios and have achieved good results. These anti-fraud algorithms either perform fraud detection from the perspective of individual users,or perform fraud detection from the perspective of topological relationship between nodes and network,or perform fraud detection by learning the graph embedded representation of nodes,which are limited in their starting perspectives and cannot perform a complete fraud detection analysis. To address the above problems,this paper designs a Kcore graph convolutional neural network anti-fraud algorithm based on the fusion of multi-source graph features. The innovation of this algorithm lies in the fact that it can efficiently mine the topological relationships at the node level in the network and the topological relationships at the global network level to build a wide-field feature system,and complete the propagation and aggregation of deep-level graph structure features through the graph convolutional neural network based on the Kcore algorithm The final result is the detection of fraud risk. Experimental results show that the method has a large improvement in the evaluation indexes compared with related machine learning algorithms and graph neural network algorithms,including a 12% improvement in the AUC value compared with LightGBM algorithm and a 6% improvement in the AUC value compared with GCN algorithm.

Keywords Machine learning,Graph representation learning,Graph neural network,Financial fraud detection

1 引言

人工智能助力的金融科技产业发展已经给社会的许多方面带来了积极的影响,企业可以更高效地组织管理资金链,个人可以更精准地完成个人金融投资与财产管理。但是层出不穷的金融欺诈也为企业和人民带来了巨大的经济损失。中国工商银行安全攻防实验室发布的《数字金融反欺诈技术应用分析报告(2021年)》表明,2019年—2022年度我国受到的欺诈损失逐年增加[1],截至2022年度,各种欺诈方式所带来的经济损失总共达7100亿元,占我国国内生产总值的0.91%。为了应对猖獗的金融欺诈攻势,多种人工智能与金融反欺诈算法被应用于实际反欺诈业务场景并取得了不错的成绩。这些算法主要可以分为基于规则引擎的反欺诈算法、基于机器学习的反欺诈算法和基于图表示学习的反欺诈算法。基于规则引擎的反欺诈算法依赖于专家经验规则的总结,具备较高的业务可解释性,在欺诈检测的同时可以给出判别欺诈风险的原因,但无法随着业务场景的快速迭代而及时迁移更新。基于机器学习的反欺诈算法依赖机器学习分类模型对海量客

基金项目:河北省重点研发计划(21310101D)
This work was supported by the National Key Research and Development Program of Hebei Province,China(21310101D).
通信作者:宋友(songyou@buaa.edu.cn)

户特征进行分析而完成欺诈风险判别，模型可以随着业务场景的迁移变换完成参数的调优，具备较高的灵活性和高效性。基于深度学习的反欺诈算法依赖人工神经网络来完成欺诈风险检测，识别准确率进一步提高，但是神经网络对于业务人员而言是完全的“黑箱”，进而缺乏足够的业务可解释性。同时，基于机器学习和深度学习的反欺诈算法对个体之间的关联信息挖掘不够深入，无法利用较深层次的网络结构特征。欺诈分子的欺诈手段也随着时间推移不断更新迭代，并且欺诈分子逐渐团伙化，欺诈手段逐渐隐蔽化，这给反欺诈任务带来了巨大的挑战。

为了更有效地挖掘利用个体之间的关联信息，本文提出了融合多源图特征的 Kcore 图卷积神经网络反欺诈算法，网络反欺诈算法主要包括特征构建算法和欺诈检测模型。Kcore 图卷积神经网络反欺诈算法对特征构建算法和欺诈检测模型都做了一定优化。在特征方面，本文利用多种随机游走算法和网络中心性分析算法来挖掘用户网络中的结构化特征。在模型方面，本文提出了基于 Kcore 子核分解的图卷积神经网络模型，从网络演进的角度挖掘深层次的网络结构信息，并通过图卷积神经网络完成邻域节点信息向目标节点信息的聚合。

2 相关工作

2.1 基于规则引擎的反欺诈算法

基于规则引擎的欺诈检测算法来源于基于规则的专家系统[2]。在行业内具备丰富反欺诈经验的专家根据过往欺诈案例总结出反欺诈规则库，搭配事实集和推理引擎即可组成规则引擎。其中事实集又名工作空间，主要用于存储用于规则匹配的事实，而推理引擎主要用来将涉及到的事实和相应的规则进行匹配。基于规则引擎的欺诈检测算法在海量大数据的业务场景中具备很高的执行效率。现有规则引擎算法主要基于 Rate 算法完成匹配[3]，将专家规则编译成 Rate 网络来与事实进行匹配。相关学者从节点索引、图算法与规则加载时机选择等方面来完成对 Rate 算法的改进和优化[4-6]。

2.2 基于机器学习的反欺诈算法

现有欺诈手段更新迭代速度较快，而基于专家经验的规则总结周期较长，随着机器学习算法的逐渐成熟，越来越多的学者将机器学习模型应用在具体的反欺诈业务场景中，并取得了不错的效果。Emekter 等[7]基于 Lending Club 的公开数据集构建了逻辑回归模型来评估信贷场景下的欺诈风险。Kruppa 等[8]使用随机森林模型、KNN 模型和 BNN 模型与逻辑回归模型进行对比，结果证明随机森林模型的反欺诈效果更好。

由于金融交易中的疑似欺诈行为可以被异常检测算法所识别，因此 Yu 等[9]提出了基于相互强化的局部异常值检测算法，该算法能够有效识别可疑欺诈行为。Robinsonu 等[10]引入隐马尔可夫模型对信用卡交易中的操作序列进行建模，并实现了面向动态交易行为场景的自动欺诈检测模型。Lucas 等[11]将多视角隐马尔可夫模型和基于专家经验的欺诈检测模型相结合，进一步提高了欺诈检测的准确率。相较于单一分类器模型，模型融合可以在一定程度上提高算法的准确率。相对于 XGBoost 算法，LightGBM 算法具备轻量化的优势，即训练速度快且模型优化容易。Wang[12]提出了基于 LightGBM 模型的银行用户信用风险分析算法，该算法不仅泛化能力快，而且稳健性较强，在相关欺诈检测的数据集上取得了不错的效果。

2.3 基于图表示学习的反欺诈算法

基于图表示学习的欺诈检测技术主要通过对网络拓扑结构的挖掘来构建图结构特征信息，以交付下游欺诈检测算法。其中基于社团划分的欺诈检测算法主要通过挖掘网络结构中的桥接节点与桥接边以判定欺诈分子与欺诈行为，在特定业务场景下效果较好，但是算法前置条件严苛，实现难度较大。基于图神经网络的欺诈检测技术主要通过将网络特征信息映射到低维向量空间，学习节点信息的隐含表示来实现欺诈检测任务。该技术的优点在于可以有效处理节点之间的非线性关系，并且低维向量空间的特征信息可以最大程度减少相关噪声的影响。因此基于图表示学习欺诈检测技术能够灵活应用于多个业务场景。Dreżewski[13]等利用 PageRank 值结合中心性度量方法表征网络结构特征，识别交易过程中具备潜在欺诈风险的用户。Akoglu 等[14]提出了基于 egonet 特征的异常检测算法。Wang 等[15]提出了一种基于全局网络结构的无监督欺诈检测算法，该算法主要从金融交易流程中提取出 3 种核心特征，分别是节点度连接性特征、边连接性特征与中心性度量特征，然后采取集成多种机器学习模型的算法进行欺诈检测。Kipf 等[16]将谱域图卷积神经网络应用于引文网络的半监督节点分类问题并取得了不错的效果。作者通过局部一阶近似性来确定了一种可缩放的卷积网络结构，并且该网络结构主要基于层向的传播规则，因此具备较好的叠加性和延展性。Xu 等[17]提出了一种图聚类算法 SCAN，SCAN 在对网络进行聚类分析的同时也完成了对桥接节点和桥接边的挖掘。Shi 等[18]提出了基于分类对抗正则器和潜在分布对齐正则器的 DR-GCN 来解决欺诈场景中的样本不平衡问题。Liu 等[19]提出 CTGCN 模型进行欺诈识别，CTGCN 能够兼具局部结构相似性和全局结构相似性。

3 融合多源图特征的 Kcore-GCN 反欺诈算法

本文提出了融合多源图特征的 Kcore-GCN 反欺诈算法，该算法主要包括多源图特征构建层和 Kcore-GCN 模型层。具体算法架构如图 1 所示。针对个体业务特征无法体现用户全部行为特征的问题，多源图特征构建算法可以在个体业务特征的基础上增加一系列衍生特征。针对普通 GCN 模型无法挖掘深层次图结构特征的问题，Kcore-GCN 模型层利用基于 Kcore 算法的图卷积神经网络模型对多源图结构特征构建层输出的特征进行分类，进而对用户网络中的个体节点完成欺诈风险判定。

3.1 多源图特征构建层

针对个体信息特征仅能代表个人行为信息，无法反映个体之间的关联信息的问题，本文提出了多源图特征构建层来挖掘节点之间隐含的关联特征信息与节点与网络之间的潜在嵌入表示。多源图特征构建层主要包括随机游走特征构建算法与网络节点中心性特征构建算法。随机游走算法通过构造随机游走序列的方式能够高效地学习节点在网络中的图嵌入表示，挖掘节点在网络中的深层向量表示，反馈节点与网络之间的特征信息。网络节点中心性特征构建算法主要挖掘网络中的节点的度中心性特征来反映网络中节点与节点之间的

关联特征。在网络分析算法中，越是趋向于网络中心位置的节点，在网络中的重要性与不可替代性越高。在风险控制反欺诈领域，活跃的节点隐含着更高的欺诈风险，因此通过网络节点中心性特征构建算法挖掘网络中每一个节点的度中心性特征可以有效反映节点的欺诈概率。

3.1.1　随机游走特征构建算法

随机游走特征构建算法主要用来研究网络中节点的低维向量表示并将该向量表示应用于下游反欺诈任务中，其核心步骤在于通过随机游走算法产生节点序列，并从该序列中挖掘节点的图嵌入表示特征。本文所采取的随机游走特征构建算法主要包括 DeepWalk[20] 算法、Node2Vec[21] 算法与 Stru2Vec[22] 算法。DeepWalk 算法首次将深度学习的思想引入图表示学习领域，能够有效地学习到节点之间的近邻相似性。Node2vec 算法相较于 DeepWalk 算法引入了有偏向的随机游走序列生成算法，可以控制随机游走序列偏向于深度优先探索的方向或偏向于宽度优先探索的方向。因此，相较于 DeepWalk 算法，Node2Vec 算法不仅能够学习到节点之间的近邻相似性，还能够学习到更深层次的节点结构相似性。但是由于传统的随机游走算法对于步长的限制，DeepWalk 无法学习到相距较远的两个节点之间的结构相似性，因此 Stru2Vec 被提出用于针对节点之间的全局结构相似性进行学习。Struc2Vec 提出层次结构来衡量节点之间的结构相似性，通过构造多层次图结构的方法来计算节点之间的相似性并生成节点向量的嵌入表示。

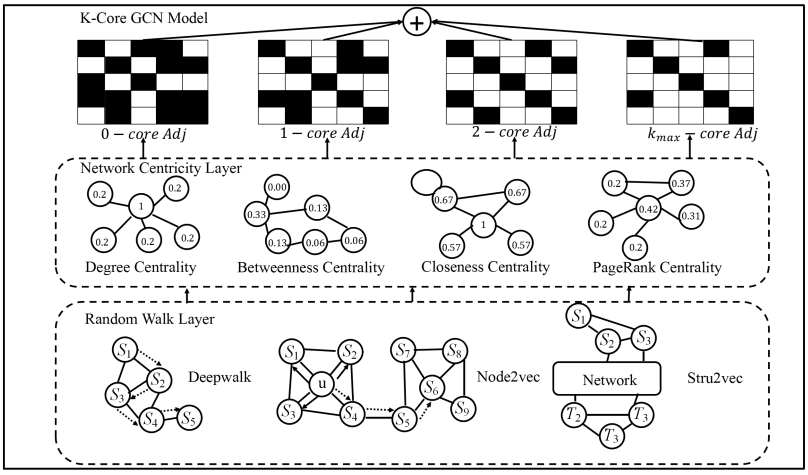

图 1　算法整体框架图

Fig. 1　Overall framework diagram of the algorithm

3.1.2　DeepWalk 算法

DeepWalk 算法主要包括两个步骤，首先通过随机游走算法完成对图中节点序列的采集，然后通过 Skip-Gram 模型获取节点的向量表示。具体步骤为，随机游走器在输入图(G)中均匀采样一个节点v_i作为根节点，进行随机游走产生长度为 t 的序列w_{v_i}，然后利用 Skip-Gram 模型对序列w_{v_i}进行学习得到节点v_i的潜在向量表示$Embedding_{v_i}$。DeepWalk 算法的目标函数为：

$$minimize_{\phi} -\log \Pr(\{v_{i-w}\cdots v_{i+w}\}\mid \phi(v_i)) \tag{1}$$

其中，$\phi(v_i)$为节点v_i的向量嵌入表示。DeepWalk 算法的优点在于具备自适应性和社团一致性。自适应性表现在对于网络中新增节点，不需要重新学习全部已有节点的向量表示，仅需学习新增节点部分即可。社团一致性表现在不同节点在不同嵌入维度的距离远近具备一致性。DeepWalk 算法生成的特征会作为 Kcore-GCN 模型训练输入特征的一部分。

3.1.3　Node2Vec 算法

Node2vec 算法针对 DeepWalk 随机游走自由度过高的问题，引入了参数 p 和 q 来平衡控制游走方向为深度优先搜索和宽度优先搜索，从而保证随机游走算法不仅能够挖掘节点之间的近邻相似性而且能够挖掘节点与网络之间的结构相似性。其中在随机游走中，当前节点选择下一个节点的转移概率为：

$$\pi_{wx}=\alpha_{p,q}(t,x)*w_{wx} \tag{2}$$

其中，$\alpha_{p,q}(t,x)$的取值由 p 和 q 来控制，决定当前节点倾向于向网络的深层次挖掘还是向当前节点的其他邻居节点处探索。Node2Vec 算法的目标函数为：

$$\max_{f}\sum_{u\in v}\left[-\log Z_u+\sum_{n_i\in N_s(u)}f(n_i)*f(u)\right] \tag{3}$$

其中，$f:V\rightarrow w^d$是从节点 V 到 d 维特征向量表示的映射函数；Z_u为节点的划分函数，可以通过负采样算法对复杂度进行优化。Node2Vec 算法生成的特征用于平衡 DeepWalk 算法生成特征游走随机性高的问题。

3.1.4　Stru2Vec 算法

由于 DeepWalk 和 Node2Vec 算法中随机游走生成的序列长度受到限制，因此无法充分获取网络中的结构特征信息，即在网络中不仅直接相连的节点具有一定的同质性，而且相距较远的两个节点之间也可能具备较高的同质性。因此 Stru2vec 算法不依赖于节点互相接近的程度来评估节点的相似性，而是通过建立一个层次结构来描述节点之间的结构相似性，而且该层次结构对于结构相似性描述的严格程度逐层递增。相似性计算依据为：

$$f_k(u,v)=f_{k-1}(u,v)+g(s(R_k(u)),s(R_k(v)),\quad k\geqslant 0 \text{ and } |R_k(u)|,|R_k(v)|\geqslant 0 \tag{4}$$

其中，$f_k(u,v)$表示节点 u 与节点 v 在考虑 $K\text{-}hop$ 领域时的结构距离，$R_k(u)$为节点 u 在 $K\text{-}hop$ 处的邻居节点集合，$s(u)$表示节点 u 在 $K\text{-}hop$ 处近邻节点的有序序列。最底层的结构相似性仅取决于节点之间的相似性，最高层的结构相似性取决于所在网络的整体相似性。在建立的层次结构中，采取多层图的有偏随机游走算法得到游走序列，进而生成节点的

嵌入表示，并将此嵌入表示应用于下游任务中。Stru2Vec 算法生成的特征用来补充 DeepWalk 算法和 Node2Vec 算法生成的特征对结构信息感知较差的问题。

3.1.5 网络中心性特征构建算法

度中心性特征构建算法旨在挖掘节点在复杂网络中的特征信息，其中主要计算的网络中心性包括：度中心性（Degree Centrality）、中介中心性（Betweenness Centrality）、接近中心性（Closeness Centrality）和特征向量中心性（Eigenvector Centrality）。

度中心性是刻画网络中节点中心性最为直观的度量标准，一个节点的度中心性数值越高，代表该节点所相连的节点数目越多，进而该节点的重要性越高。度中心性的计算式如式(5)所示：

$$\boldsymbol{DC}_i=\frac{k_i}{N-1} \tag{5}$$

其中，k_i 表示节点 i 的度；$N-1$ 表示节点 i 的最大度值，即节点 i 与其他节点都相连的边的数量。

节点介数是指网络中通过目标节点的最短路径条数，中介中心性主要是基于最短路径的概念来刻画节点的中心性，其计算式如式(6)所示：

$$\boldsymbol{BC}_i=\sum_{s\neq i\neq t}\frac{n_{st}^i}{g_{st}} \tag{6}$$

其中，g_{st} 表示连接节点 s 和 t 的最短路径的数量，n_{st}^i 表示连接节点 s 和 t 且经过节点 i 的最短路径的数量。

接近中心性刻画了网络中目标节点与其他节点之间的接近程度，即计算目标节点与其他所有节点的距离总和，该距离值越小，代表目标节点与其他节点的“距离”越接近。其计算式如式(7)所示：

$$\boldsymbol{CC}_i=\frac{n}{\sum_{j=1}^{N}d_{ij}} \tag{7}$$

其中，d_{ij} 表示节点 i 和 j 的最短距离。

特征向量中心性是基于与高分值节点的连接比与低分值节点的连接对于目标节点的分值贡献更大的概念来讨论节点在网络中的重要性程度。对于节点 V_i，其特征向量中心性的计算公式为：

$$\boldsymbol{C}_e(V_i)=\frac{1}{\lambda}\sum_{j=1}^{N}\boldsymbol{A}_{i,j}\boldsymbol{C}_e(V_j) \tag{8}$$

其中，$\boldsymbol{C}_e\in R^N$ 是一个包含所有节点的特征向量中心性的向量，$\boldsymbol{C}_e$ 是其对应的特征值，$\boldsymbol{A}$ 为邻接矩阵。

PageRank 中心度是基于 PageRank 算法得到的衡量节点中心性的数值。当 PageRank 算法中的节点之间的转移概率趋于平稳时，即可得到收敛后的概率矩阵 $\boldsymbol{R}$，$\boldsymbol{R}$ 的各个分量为每一个节点 v_i 的 PageRank 中心度 $PR(v_i)$，即：

$$\boldsymbol{R}=\begin{bmatrix}PR(v_1)\\PR(v_2)\\\cdots\\PR(v_n)\end{bmatrix} \tag{9}$$

网络中心性特征构建算法所构建的特征能够更好地表征网络中的结构特征信息，因此同样将其作为 Kcore-GCN 模型训练输入特征的一部分。

3.2 Kcore-GCN 模型层

3.2.1 图卷积神经网络算法（GCN）

图卷积神经网络算法能够学习到网络中节点的低维度向量嵌入表示。相较于上文提到的随机游走类算法，图卷积神经网络算法能够端到端地完成下游训练任务，如节点分类与链接预测等。其中 GCN 的核心公式为：

$$\boldsymbol{H}^{l+1}=\sigma(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{H}^l\boldsymbol{W}) \tag{10}$$

其中，$\boldsymbol{H}^l$ 为网络中节点在 l 层的节点嵌入表示，$\boldsymbol{A}$ 为带有自环边的图的邻接矩阵。$\boldsymbol{D}$ 为图中节点的度矩阵，表征图中节点的度数，由式(11)计算得到。

$$D_{ii}=\sum_{j=1}^{N}A_{ij} \tag{11}$$

一层图卷积神经网络相当于将目标节点的一阶邻居节点信息聚合到目标节点上，并且在常规的 GCN 算法中，该聚合为均值聚合，即不区分目标节点的不同邻居节点的特质性。另一方面，在常规图卷积神经网络算法中，图卷积层数不会很大，一般为 2－3，因此无法学习到网络中深层次的结构化信息。并且如果简单地堆叠图卷积层，会导致网络中节点的特征表示趋于一致性，一般称作过平滑问题（Over Smoothing）。

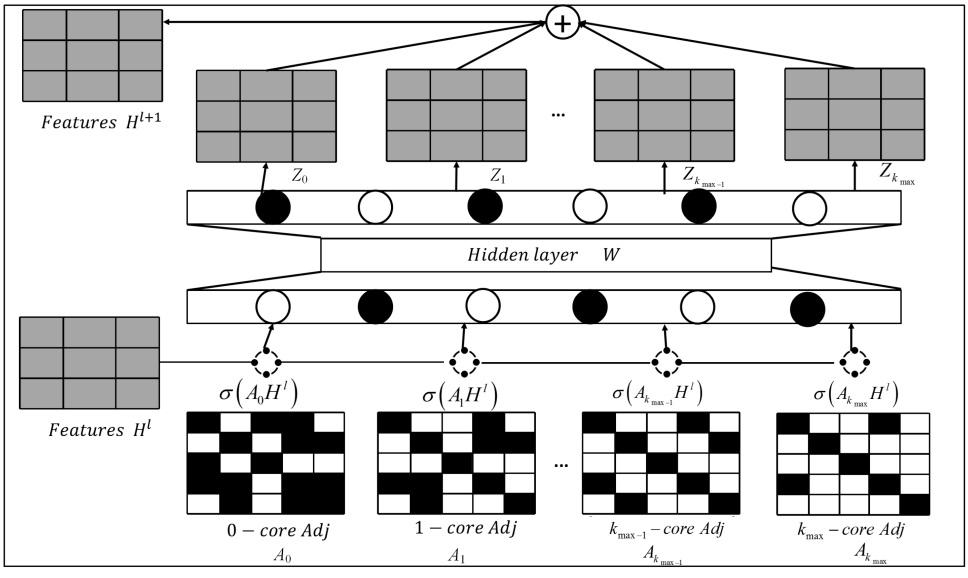

图 2 Kcore-GCN 模型结构图

Fig. 2 Structure of Kcore-GCN model

3.2.2　Kcore 算法

Kcore[23]算法是一种子图挖掘算法，通常被用来在网络中挖掘出一个具备 K 核心度的子图结构，并且该子图具备较高的核心地位。随着核心度数目的增长，K 核心度子图的规模会缩减，但是子图的重要性会越来越高。下面给出 K 核心度子图的定义以及 Kcore 算法的核心流程。

K 核心度子图：图中所有节点至少具备 k 的度数且所有节点都至少与该子图中的 k 个其他节点相连。

算法 1　Kcore

输入：图 G，核心度 K

输出：符合定义的 K 核子图

1. 计算图 G 中各节点的度数
2. For 图中的每个节点 do
3. 删除度数小于 K 的顶点
4. Endfor
5. 计算图 G 中各节点的度数
6. For 图中的每个节点 do
7. 删除度数小于 K 的顶点
8. Endfor

可以证明输出的图 G 即是符合定义的 K 核子图，同时对于图 G 的 K-core 序列$\{G_0,G_1,G_2,\cdots,G_{k_{\max}}\}$，有如下所示的递归包含关系：

$$G=G_0\supseteq G_1\supseteq G_2\supseteq\cdots G_{k_{\max}} \tag{12}$$

另一方面，由于 K 核心度挖掘算法的时间复杂度仅与图 G 中节点数呈线性相关性，因此 K 核心度挖掘算法可以被应用于大规模网络性质挖掘。

3.2.3　K 核图卷积神经网络模型

K 核图卷积神经网络模型(Kcore-GCN)在常规的图卷积神经网络模型(GCN)的基础上增加了前置的 K 核心度矩阵分解层，用于解决图卷积神经网络模型在节点级别和网络级别的问题。其中节点级别的问题是目标节点聚合邻居节点信息的时候不做区分处理，而 Kcore-GCN 模型先做了 K 核心度分解算法再进行聚合。根据 Kcore 算法的分解过程可知，核心度越高的节点在被目标节点聚合时权重越高。网络级别的问题是指常规的图卷积神经网络算法无法挖掘深层次的网络结构化信息，而 Kcore 子图序列符合网络演进的渐近性，即核心度越高的子图结构越可以代表网络的深层次的结构信息。结合算法流程图可知，对于输入的图结构化数据 $G=(V,E)$，可进行 Kcore 分解算法，得到图 G 的 Kcore 序列$\{G_0,G_1,G_2,\cdots,G_{k_{\max}}\}$。可以求得对应 Kcore 核心度子图的邻接矩阵序列。

$$\{\boldsymbol{A}_0,\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_{k_{\max}}\} \tag{13}$$

对于第 l 层的节点图嵌入向量H^l，我们定义特征聚合函数 $Z=f(x)$，那么对于 Kcore 子图G_k则有：

$$Z_k=(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}_k\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{H}^l\boldsymbol{W}) \tag{14}$$

其中，σ 为激活函数，在本实验中选取为 Relu 激活函数。

$$\boldsymbol{H}^{l+1}=\sum_{i=0}^{k_{\max}}Z_i \tag{15}$$

在模型输出时将采用相关聚合函数对所有 Kcore 子图G_k完成聚合，其中聚合函数可以选取求和、均值、RNN 等聚合函数。

4　实验结果与分析

本文实验采取 Python3.6 及 Pytorch1.1 深度学习框架搭建模型，并使用 AUC 值作为评价指标，在真实金融业务数据集上进行实验探究和验证。在实验阶段，将数据集的 80% 作为训练集，20%作为测试集。为保证实验结果的一般性，实验的 AUC 值得采用 5 折交叉验证的均值。

4.1　数据集介绍

本实验采取公开数据集 YelpChi 评论欺诈数据集。Yelp 是美国最大点评网站，具有较大的顾客群体和社会影响力。其中 YelpChi 数据集是基于 Yelp 的一个行为图数据集，数据集中个体特征与图结构化特征以稀疏矩阵的形式进行存储。该数据集被广泛应用于金融风控、欺诈检测、反洗钱等研究任务中。YelpChi 欺诈数据集执行欺诈检测任务，该任务本质上是一个二元分类任务，即是否存在潜在欺诈风险。YelpChi 提取了 32 个手工特征作为节点的原始特征，节点之间的关系根据业务属性被划分为 3 类。其中 YelpChi 行为图中包括 45954 个点、144584 条边，属于较大规模的图网络。

4.2　对比算法介绍

(1)逻辑回归算法：逻辑回归算法被广泛应用于信贷欺诈风险预测中，例如知名的 FICO 评分就是基于逻辑回归算法建立的。

(2)决策树算法：决策树算法主要利用归纳分析算法总结生成可读性较高的分类规则和决策树，然后对测试数据利用生成的决策树完成分类任务。

(3)K 近邻算法：K 近邻算法作为一种被广泛应用的机器学习算法，具备训练难度低、训练时间少、可以完成海量数据的分类问题等优点。

(4)XgBoost：XGBoost 是基于集成学习的算法框架，是对 GBDT 算法的高效系统化的实现。Xgboost 使用牛顿法求解损失函数极值来优化梯度提升算法。

(5)LightGBM：LightGBM 是微软提出的基于集成学习的算法框架，其原理与 XgBoost 相似，但 LightGBM 针对训练过程做出了优化，因此相较于 XgBoost 有模型轻量级和训练速度快等优点。

4.3　评价指标

本文选取 AUC 值作为实验的评价指标。

4.4　实验分析

4.4.1　随机游走特征构建算法有效性验证

为验证随机游走特征构建算法的有效性，本实验选取基础特征与基础特征和 3 种不同的随机游走算法构建的特征相组合的特征在 5 种分类算法的情况下进行比较，其中不同分类器算法的参数均经过多次实验确定为最佳表现的参数。为保证不同随机游走算法构建特征的有效对比，3 种随机算法所构建的特征均为 64 维。实验效果表明，相较于只使用基础特征，基础特征与随机算法构建特征的组合能够有效提高欺诈检测的 AUC 值，根据 5 种分类算法的表现可知，DeepWalk 算法 AUC 值的提升幅度为 0.01～0.094，Node2Vec 算法 AUC 值的提升幅度为 0～0.051，Stru2Vec 算法的提升幅度为0.003～0.096。3 种随机游走特征构建算法中 Node2Vec 算法较为一般，DeepWalk 算法和 Stru2vec 算法均有不俗的提升。5 种分类算法中 XgBoost 和 LightGBM 表现优异，并且 LightGBM 在基础特征和 Stru2vec 算法的组合特征上达到了本次实验最好的 AUC 值，为 0.764。

表 1 3 种随机游走算法对比结果

Table 1 Comparison of three random walk algorithms

模型	Basic 特征	Basic 特征＋DeepWalk	Basic 特征＋Node2vec	Basic 特征＋Stru2vec
LR	0.501	0.508	0.501	0.504
DecisionTree	0.525	0.568	0.527	0.549
KNeighbors	0.536	0.630	0.570	0.600
XgBoost	**0.668**	0.744	0.688	**0.764**
LightGBM	0.665	**0.758**	**0.716**	0.761

4.4.2 网络中心性算法有效性验证(Deepwalk)

为验证网络中心性特征构建算法的有效性，本实验选取基础特征，基础特征和 DeepWalk 算法构建特征的组合特征，基础特征和 DeepWalk 算法和网络中心性特征的组合特征在 5 种分类器算法的情况下进行对比。实验表明网络中心性特征构建算法可以在随机游走特征构建算法的基础上进一步提升欺诈检测的 AUC 值，其中决策树算法、K 近邻算法和 LightGBM 算法的提升程度在 0.005 左右，而 XgBoost 算法在 DeepWalk 特征构建算法的基础上提升了 0.022，进而达到本次实验最好的 AUC 值，为 0.766。

表 2 DeepWalk 算法和网络中心性算法组合效果对比

Table 2 Comparison of the combination effects of DeepWalk algorithm and network centrality algorithm

模型	Basic 特征	Basic 特征＋DeepWalk	Basic 特征＋DeepWalk＋接近中心性
LR	0.501	0.508	0.506
DecisionTree	0.525	0.568	0.584
KNeighbors	0.536	0.630	0.635
XgBoost	**0.668**	0.744	**0.766**
LightGBM	0.665	**0.758**	0.761

4.4.3 网络中心性算法有效性验证(Stru2Vec)

为验证网络中心性特征构建算法的有效性，本实验选取基础特征，基础特征和 Stru2vec 算法构建特征的组合特征，基础特征和 Stru2vec 算法和网络中心性特征的组合特征在 5 种分类器算法的情况下进行对比。可以发现网络中心性特征构建算法在 Stru2vec 算法的基础上 AUC 值的提升幅度较低，说明网络中心性特征构建算法和不同随机游走算法的适配程度不同。其中 XgBoost 算法在 Stru2vec 特征构建算法的基础上提升了 0.003，达到了本次实验最好的 AUC 值，为 0.767。

表 3 Stru2vec 算法和网络中心性算法组合效果对比

Table 3 Comparison of the combination effects of Stru2Vec algorithm and network centrality algorithm

模型	Basic 特征	Basic 特征＋stru2Vec	Basic 特征＋stru2vec＋接近中心性
LR	0.501	0.504	0.503
DecisionTree	0.525	0.549	0.550
KNeighbors	0.536	0.600	0.601
XgBoost	**0.668**	**0.764**	**0.767**
LightGBM	0.665	0.761	0.762

4.4.4 Kcore-GCN 有效性验证

为证明 Kcore-GCN 模型的有效性，本实验选取基础特征，基础特征和 DeepWalk 算法构建特征和网络中心性特征的组合，基础特征和 Stru2vec 算法构建特征和网络中心性特征的组合这 3 种特征作为特征层面的对照。在模型选取方面，除去上面实验涉及的 5 种机器学习分类模型，同时将常规的 GCN 和 Kcore-GCN 模型进行对比。可以发现 Kcore-GCN 模型分类效果优于 LightGBM 模型、XgBoost 模型和常规的 GCN 模型。其中 Kcore-GCN 模型在 3 种特征的 AUC 值较 GCN 模型均有 0.01 的提升幅度。

表 4 Kcore-GCN 模型有效性验证

Table 4 Validation of Kcore GCN model

模型	Basic 特征	Basic 特征＋DeepWal＋接近中心	Basic 特征＋Stru2vec＋接近中心性
LR	0.501	0.506	0.503
DecisionTree	0.525	0.584	0.550
KNeighbors	0.536	0.635	0.601
XgBoost	0.668	0.766	0.767
LightGBM	0.665	0.761	0.762
GCN	0.727	0.759	0.780
Kcore-GCN	**0.737**	**0.767**	**0.789**

结束语 本文针对现有反欺诈算法难以充分挖掘网络结构中节点之间的关联信息和网络结构深层次信息的问题，提出了基于融合多源图特征的 Kcore-GCN 算法，该算法在特征层面对比分析了 3 种随机游走特征构建算法和基于网络中心性构建的算法特征。在模型层面，相较于普通的 GCN 模型，提出了基于 Kcore 子核挖掘算法的 Kcore-GCN 模型。实验表明融合多源特征相较于单一个体特征在多种分类器模型下 AUC 值均有提升，并且 Kcore-GCN 模型相较于多种机器学习算法和普通 GCN 模型表现更好。其中融合多源图特征的 Kcore-GCN 算法较 LightGBM 算法有 12%的 AUC 值提升，较 GCN 算法有 6%的 AUC 值提升。后续可以利用大规模随机游走算法 LINE 等和基于注意力机制的图卷积神经网络来优化欺诈检测的准确度。

参 考 文 献

[1] 中国工商银行金融科技研究院安全攻防实验室，中国信息通信研究院云计算与大数据研究所. 数字金融反欺诈技术应用分析报告[EB/OL]. https://www.freebuf.com/articles/paper/325372.html.

[2] GU X D. Research and application of large-scale rule reasoning engine based on rule algorithm[D]. Nanjing University, 2013.

[3] FORGY C L. Rete: A fast algorithm for the many pattern/many object pattern match problem[J]. Artificial Intelligence, 1982, 19(1): 17-37.

[4] LIU D, GU T, XUE J P. Rule Engine Based onimprovement Rete algorithm[C]// The 2010 International Conference on Apperceiving Computing and Intelligence Analysis Proceeding. Chengdu, 2010: 346-349.

[5] SUN Y, WU T Y, ZHAO G, et al. Efficient Rule Engine for Smart Building Systems[C]// IEEE Transactions on Computers. 2015: 1658-1669.

[6] CHATTOPADHYAY S, BANERJEE A, BANERJEE N. A Scalable Rule Engine Architecture for Service Execution Frameworks[C]// 2016 IEEE International Conference on Services Computing(SCC). San Francisco, CA, 2016: 689-696.

[7] TU E, LU J. Evaluating credit risk and loan performance in online Peer-to-Peer(P2P) lending[J]. Applied Economics, 2015, 47(1): 54-70.

[8] KRUPPA J, SCHWARZ J, ARMINGER G, et al. Consumer Credit Risk: Individual Probability Estimates Using Machine

Learning[J]. Expert Systems with Applications，2013，40(13)：5125-5131.

[9] YU J X，QIAN W N，LU H J，et al. Finding centric local outliers in categorical/numerical spaces[J]. Knowledge and Information Systems，2006，9(3)：309-338.

[10] ROBINSON W N，ARIA A. Sequential fraud detection for prepaid cards using hidden Markov model divergence[J]. Expert Systems with Applications，2018，91：235-251.

[11] LUCAS Y，PORTIER P E，LAPORTE L，et al. Towards automatedfeature engineering for credit card fraud detection using multi-perspective HMMs[J]. Future Generation Computer Systems-the International Journal of Escience，2020，102：393-402.

[12] WANG J C. Rredit Risk Analysis of Bank Users Based on Machine Learning Algorithm [D]. Tianjin：Nankai University，2021.

[13] DREŻEWSKI R，SEPIELAK J，FILIPKOWSKI W. The application of social network analysis algorithms in a system supporting money laundering detection[J]. Information Sciences，2015，295：18-32.

[14] AKOGLU L，MCGLOHON M，FALOUTSOS C. Oddball：Spotting anomalies in weighted graphs[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin，Heidelberg：Springer，2010：410-421.

[15] WANG K，CHEN D. Graph structure based anomaly behavior detection[C]// 2nd International Conference on Computer Engineering，Information Science & Application Technology(ICCIA 2017). Atlantis Press，2016：531-538.

[16] KIPF T N，WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv：1609. 02907，2016.

[17] XU X，YURUK N，FENG Z，et al. Scan：a structural clustering algorithm for networks [C] // Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2007：824-833.

[18] SHI M，YANG Y F，ZHU X Q，et al. MultiClass Imbalanced Graph Convolutional Network Learning[C]// IJCAI. 2020.

[19] LIU J，XU C，YIN C，et al. K-core based temporal graph convolutional network for dynamic graphs[J]. IEEE Transactions on Knowledge and Data Engineering，2020，34(8)：3841-3853.

[20] 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. 2014：701-710.

[21] 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. 2016：855-864.

[22] LI Y，LIU X，WANG C. Research on Link Prediction under the Structural Features of Attention Stream Network [C] // 2021 IEEE Asia-Pacific Conference on Image Processing，Electronics and Computers(IPEC). IEEE，2021：148-154.

[23] KONG Y X，SHI G Y，WU R J，et al. k-core：Theories and applications[J]. Physics Report，2019，832：1-32.

LIU Wei，born in 1998，master. His main research interests include data mining and graph neural networks.

SONG You，born in 1973，Ph.D，professor. His main reaserch interests include software engineering，big data analysis，technology finance，and so on.