2025 年 第 25 卷 第 20 期
2025, 25(20): 08604-11

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2405389
引用格式:张陶, 廖彬, 于炯. 基于图嵌入与集成分类算法的内容特征缺失网页多分类预测方法[J]. 科学技术与工程, 2025, 25(20): 8604-8614.
Zhang Tao, Liao Bin, Yu Jiong, et al. Multi-classification prediction of web pages with missing content features based on graph embedding and ensemble classification algorithm[J]. Science Technology and Engineering, 2025, 25(20): 8604-8614.

基于图嵌入与集成分类算法的内容特征缺失网页多分类预测方法

张陶[1,2], 廖彬[3*], 于炯[2]
(1. 贵州中医药大学信息工程学院, 贵阳 550025; 2. 新疆大学信息科学与工程学院, 乌鲁木齐 830046;
3. 贵州财经大学大数据统计学院, 贵阳 550025)

摘 要 由于噪声(如广告)、权限不足、隐私保护或恶意伪装等原因,造成大量网页显式内容特征不能被及时、全面的获取。在此背景下,为解决在网页内容特征严重缺失情况下如何对网页有效分类的问题,提出一种基于图嵌入与集成分类算法 XGBoost(extreme gradient boosting)利用网页链接网络中隐含关系特征进行网页多分类的方法。首先,利用网页及网页间的超链接关系,构造出网页链接网络;然后,通过图嵌入模型抽取节点(网页)在链接网络中的隐含关系特征;其次,提取节点的集聚系数、PageRank 值等统计学结构特征,共同构成节点的稠密特征向量;最后,利用基于 XGBoost 等集成学习模型构建节点分类预测模型对网页进行分类预测。在真实维基百科网页链接数据集上的实验结果表明:在完全缺乏网页显式内容特征情况下,所提出的 Struct2Vec * +XGBoost 组合方案实现了良好的网页分类效果,在准确率、精准率、查全率及 F_1 值 4 项指标上分别达到 0.987 5、0.965 9、0.971 3 和 0.964 1。

关键词 内容特征缺失; 图嵌入; 网页链接网络; 网页多分类
中图法分类号 TP393; **文献标志码** A

Multi-Classification Prediction of Web Pages with Missing Content Features Based on Graph Embedding and Ensemble Classification Algorithm

ZHANG Tao[1,2], LIAO Bin[3*], YU Jiong[2]
(1. College of Information Engineering, Guizhou University of Traditional Chinese Medicine, Guiyang 550025, China;
2. School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China;
3. College of Big Data Statistics, Guizhou University of Finance and Economics, Guiyang 550025, China)

[**Abstract**] Explicit content features of webpages are often unavailable due to distractions such as commercials, insufficient permissions, privacy protection, or deceptive disguises. To address the challenge of classifying webpages with severe content feature deficiency, a method combining graph embedding and extreme gradient boosting(XGBoost) was proposed. This method leveraged implicit relational features in webpage hyperlink networks for multi-classification. Firstly, a hyperlink network was constructed using relationships between webpages. Then, node features were extracted using graph embedding models, and statistical structural features such as clustering coefficients and PageRank values were concatenated to form dense feature vectors. Finally, ensemble learning models, including XGBoost, were trained to classify webpages for prediction. Experiments on a real Wikipedia dataset show that the Struct2Vec * + XGBoost approach achieves excellent classification results, with accuracy, precision, recall, and F_1-score metrics reaching 0.987 5, 0.965 9, 0.971 3, and 0.964 1, respectively. These results are superior to those of comparison models. The findings demonstrate the effectiveness of using implicit link-based features for webpage classification in scenarios with content feature deficiency.

[**Keywords**] missing content features; graph embedding; webpage hyperlink network; webpage multi-classification

互联网的迅速发展带来了海量网页数据,这些数据中蕴含着巨大价值,对学者而言,如何有效管理和挖掘这些数据已成为重要课题。网页分类作为网页挖掘的基础,对于识别欺诈网页、排序网页重要性等应用至关重要[1-2]。现有绝大部分网页分类主要基于预设主题,通过从网页的文本、网页结构和超链接 URL 等显式信息提取特征,并运用多种机器学习模型,如逻辑回归(logistic regression,

收稿日期: 2024-07-17; **修订日期**: 2025-04-12
基金项目: 国家自然科学基金(61562078);贵州中医药大学博士启动基金([2024]07 号)
第一作者: 张陶(1988—),女,汉族,安徽阜阳人,博士,讲师。研究方向:机器学习、数据挖掘与复杂网络分析。E-mail:zt59921661@126.com。
* **通信作者**: 廖彬(1986—),男,汉族,四川内江人,博士,副教授。研究方向:机器学习、数据挖掘及大数据计算模型等。E-mail:liaobin665@163.com。

LR)、支持向量机(support vector machine, SVM)、随机森林(random forest,RF)和神经网络等进行类别判定。然而,网络环境日益复杂,噪声(如广告)、权限限制、隐私保护和恶意伪装等因素导致大量网页的显式特征难以及时全面地获取。在这种情况下,使得传统网页分类方法难以有效分类网页。

为应对网页显式内容特征缺失的挑战,提出一种新思路:将网页分类问题转化为基于网页链接网络的节点分类问题。网页链接网络由网页间的超链接构成,每个网页对应一个节点,节点属性代表网页内容,节点间的边表示链接关系。因此,即使在网页属性信息不完整的情况下,也能利用网页间的链接关系和网络拓扑结构进行分类,将网页视为无属性图中的节点,通过节点间的链接关系和网络拓扑结构进行分类。

基于此,现利用网页及网页间的超链接关系,构造出网页链接网络,并在此基础上利用 5 种主流图嵌入模型获取节点(网页)在网页链接网络中的隐含关系特征,得到每个节点的图嵌入特征向量。提取出节点的度、入度、出度、集群系数、网页重要性及 PageRank 值等多维统计学结构特征,与图嵌入特征拼接成为节点组合特征;在此基础上利用XGBoost训练出最佳多分类模型进行网页分类。通过在真实维基百科网络数据集进行实证,探索网页链接网络中的隐含关系特征对网页分类的影响,为后续相关研究提供参考。

1　相关研究

1.1　传统网页分类方法

针对网页分类问题,研究者们提出了多种基于网页显式特征的方法,包括网页文本内容特征(如标题、作者、摘要、主题内容等)和结构特征(如页面布局、HTML 标签结构等)。基于文本内容的方法[3-5]起源于文本分类,通过上下文分析、语义分析和摘要技术提取网页文本信息,然后再应用文本分类算法进行网页分类。但由于网页的半结构化特性,如广告等噪声会干扰有效文本信息的提取,导致仅依赖文本内容特征的分类效果不佳。相比之下,基于网页自身结构特征的方法[6-8]结合了网页文本特征和结构特征,通过传统机器学习算法,有效提升了网页分类效果。然而,上述两类方法的准确性受限于网页显式内容特征的提取,尤其在面对以图片为主、内容隐藏或受权限限制的网页时。提取 URL 信息以标识网页功能或主题提供了新的分类方法,但传统基于 URL 的方法[9-11]主要依赖于表面的语法或结构信息,未能充分利用网页间的隐含关系特征,导致分类效果受限。

1.2　基于图嵌入的节点分类方法

随着深度学习技术的进步,图嵌入[12-13]作为一种新兴的表示学习方法,在多个领域显示出巨大潜力。图嵌入技术将图中的节点(如网页)映射到低维向量空间,保留节点间的结构信息和关系,如相似性、链接模式和社区结构等,这些是传统内容特征难以捕获的。同时这种低维向量表示便于用于机器学习任务,如节点分类。如 DeepWalk[14]、Node2Vec[15]等图嵌入模型,首先在网络上以某种策略随机游走,捕获网络结构信息并学习节点,然后用于下游任务。Cai 等[16]提出了一种针对无属性图分类任务的基线方法,利用 Local Degree Profile 挖掘节点结构特征,利用 SVM 进行节点分类,但在 COLLAB 等数据集上的精度基本低于 80%。Rozemberczki 等[17]使用 DeepWalk, Role2Vec 学习节点特征向量,在利用逻辑回归训练分类预测。文献[18]提出的 BANE(binarized attributed network embedding)模型在复杂网络节点特征提取上,相较于 DeepWalk 等模型,F_1 指标提升了约 3%。

尽管图神经网络[19]端到端模式在隐含特征学习与分类上具有协同优势;但其性能受限于节点属性的完整性。相比之下,图嵌入算法如 DeepWalk、Node2Vec 等对节点属性的完整性几乎没有要求。因此,提出一种基于图嵌入与集成算法 XGBoost 的网页分类方法。区别在于:①区别于传统基于网页显式内容特征的分类方法,所提方法不依赖于网页的显式内容特征,而是利用网页在链接网络中的隐含关系特征进行分类,为无法获取显式内容特征的网页提供了分类的新途径;②不同于文献[16-17],采用混合特征提取策略,除了图嵌入特征外,还提取节点的统计学特征,如连接性、聚集性、重要性和中心性等,以更全面地描述网页间的关系特性,提高分类准确性;③以往的研究主要集中于生成节点嵌入特征,而在分类模型的选择上较为简单(如 SVM 模型[16,18]、逻辑回归模型[17]等),缺少对更优分类模型的探索和尝试。将 XGBoost 等集成学习模型应用到网页链接网络节点分类任务中,并通过实验验证适配集成分类模型的有效性。

2　模型设计

2.1　问题定义

定义 1　基于网页链接数据构建的网页链接网络 $G=(V,E,\Psi,\Omega)$,其中,V 为所有节点(网页)的集合,若网页总数为 n,则 $|V|=n$,第 i 个网页 $v_i \in V(1 \leqslant i \leqslant n)$;$E$ 为所有边(网页链接)的集合,若链接总数为 m,则 $|E|=m$,两个网页间的链接 e_{ij} =

$e_{dge}(v_i,v_j) \in E(1 \leqslant i \leqslant n,1 \leqslant j \leqslant n)$,其中 $e_{dge}(v_i,v_j)$ 为网页 v_i 和网页 v_j 间的链接(边);Ψ 为网页属性特征集合,节点 v_i 的属性特征为 $\Psi(v_i)$;Ω 为网页链接关系属性特征集合,边 e_{ij} 的具体属性特征为 $\Omega(e_{ij})$。当网页显式内容信息难以提取,导致节点属性特征缺失,即 Ψ 和 Ω 为空时,网页链接网络 G 为无属性图。

定义 2 无属性网页链接网络节点分类模型是,在无属性的网页链接网络 G 中,部分节点(网页)附有类别标签,而其他节点的标签未知,该任务旨在利用已知标签节点的连接关系和网络拓扑结构信息来训练模型,以便预测未标记网页的类别标签。设 Y 为网页标签集合,其中,$y_i \in Y(1 \leqslant i \leqslant n)$ 为节点(网页)v_i 所对应的类别标签,$y \in \{c_1,c_2,\cdots,c_k\}$ 为类别标签的总数。当 $k=2$ 时,表示网页二分类问题,当 $k>2$ 时,表示网页多分类问题。

定义 3 损失函数。在给定网页链接网络 G 和网页标签集合 $y_i \in Y(1 \leqslant i \leqslant n)$ 的情况下,分类模型的目标是最小化损失函数 L,其表达式为

$$L = \sum_{i=1}^{n} l[y_i, f(v_i)] \tag{1}$$

式(1)中:y_i 为节点 v_i 的真实标签;$f(v_i)$ 为网页分类模型预测的标签;l 为损失函数,用于衡量预测标签与真实标签之间的差异;n 为网页链接网络中所有网页的总数。

2.2 提取网页隐含关系特征

为提取网页链接网络中的深层隐含关系特征。通过 DeepWalk、Node2Vec 等图嵌入技术,能够捕获网页间的复杂隐含关系,并将高维网络数据降维为低维嵌入向量,同时保留节点间的隐含关系特征。图嵌入映射函数可表示为

$$f(X) \rightarrow \boldsymbol{M}_{n\times s} \tag{2}$$

式(2)中:在图嵌入模型中,X 为输入网络,即网页链接网络 G;$\boldsymbol{M}_{n\times s}$ 为训练结果,是一个 $n \times s$ 的矩阵,其中,n 为网页链接网络中网页的数量,s(本文默认为 128)为节点嵌入向量的维度。

这些向量综合了节点的局部特征以及其在整个网络中的位置与角色。

为捕捉节点在网页链接网络中的结构、位置、重要性等特征,提取包括度(degree,记为 d_{eg})、入度(in_degree,记为 i_{ndeg})、出度(out_degree,记为 o_{utdeg})和 PageRank(记为 P_R)的四维统计特征,以及集群系数(clustering coefficient,记为 c_{lu})来衡量节点的局部邻域拓扑特征,并将这些特征与节点嵌入向量结合,形成节点的特征向量。

将网页链接网络 G 中每个节点的统计特征记为 $\boldsymbol{A}_{n\times 5}$,其中任意节点 v_i 的 $\boldsymbol{A}(v_i)$ 可表示为

$$\boldsymbol{A}(v_i) = [d_{eg}(v_i), i_{ndeg}(v_i), o_{utdeg}(v_i), c_{lu}(v_i), p_g(v_i)] \tag{3}$$

式(3)中:节点 v_i 的集聚系数通过式(4)计算。

$$c_{lu}(v_i) = \frac{|\{e_{ij}\}|}{d_{eg}(v_i)[d_{eg}(v_i)-1]} \tag{4}$$

网页链接网络 G 中,节点 v_i 的 P_R(PageRank)值通过式(5)计算。

$$P_R(v_i) = \alpha \sum_{v_i \in o_{ut}(v_i)} \frac{P_R(v_i)}{|o_{utdeg}(v_i)|} + \frac{1-\alpha}{n} \tag{5}$$

式(5)中:$o_{ut}(v_i)$ 为节点 v_i 的所有出链节点集合;$|o_{utdeg}(v_i)|$ 为节点 v_i 的出链个数;α 为阻尼系数,α 取值为 0.85。

将网页链接网络 G 的节点嵌入特征矩阵 $\boldsymbol{M}_{n\times s}$、节点统计结构特征矩阵 $\boldsymbol{A}_{n\times 5}$ 和节点标签集 Y 按节点 ID 编号关联,形成最终训练数据集 D,可表示为

$$D = \{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\} \tag{6}$$

式(6)中:$|D| = |V| = n$ 为所有节点的数目;$x_i = (x_i^1, x_i^2, \cdots, x_i^{s+5})$ 为节点 v_i 最终的特征向量,其中 $s+5$ 为节点特征的维度;$y_i \in \{c_1,c_2,\cdots,c_k\}$ 为网页类别标签的取值范围。

2.3 训练 XGBoost 分类模型

XGBoost 是一种基于梯度提升决策树的集成学习方法,由 Chen 等[20]提出,因其出色的防过拟合能力和泛化性能而受到广泛认可。利用 3.2 节中挖掘出的网页隐含关系特征,训练一个性能可靠的 XGBoost模型。该模型的目标是提高网页链接网络中网页分类的精确度。

3 实验设计及分析

3.1 实验环境配置及数据集信息

本实验环境配置如下:操作系统为 ubuntu 18,Python 版本 3.7,tensorflow 版本 1.14,networkX 版本 2.2,scikit-learn 版本 0.23.1。硬件配置包括 core i7-8750h 2.20 GHz CPU 和 16 GB RAM。实验数据及代码已开源,网址为:https://gitee.com/zhang-tao66/wiki_networks_classification。

构建网页链接网络[21],采取以下步骤。

步骤 1 使用 Python 爬虫从维基百科(http://en.wikipedia.org/)上爬取工程、技术与应用科学大类下共计 17 个类别(标签)的网页,包括交通、建筑学、土木工程、电气工程等多个学科。

步骤 2 将爬取到的网页之间的链接关系通过图的边相互关联,并使用 networkX 构建网页链接网络。

步骤 3 将网页的第一个分类标签作为该内容的标签字段。

步骤 4　最终构建出的网页链接网络节点规模为 2 405 个,边规模为 17 981 条,节点的平均度值为 13.74,平均入度与出度均为 6.87,节点平均聚集系数为 0.323 8,默认参数下 PageRank 计算均值为 0.000 415 8,图密度为 0.002 858。

网络节点的度、入度和出度分布如图 1 所示,均呈现正偏和厚尾特征,具体偏度与峰度值分别为(5.561, 58.095)、(8.623, 126.606)和(2.285, 7.932)。节点聚类系数和 PageRank 值的分布如图 2 所示,聚类系数偏度和峰度为(0.711, −0.385),分布较均匀;PageRank 值偏度和峰度为(7.124, 75.294),表现出明显的正偏和厚尾特征,超 90% 节点 PageRank 值集中在[0,0.003]区间。

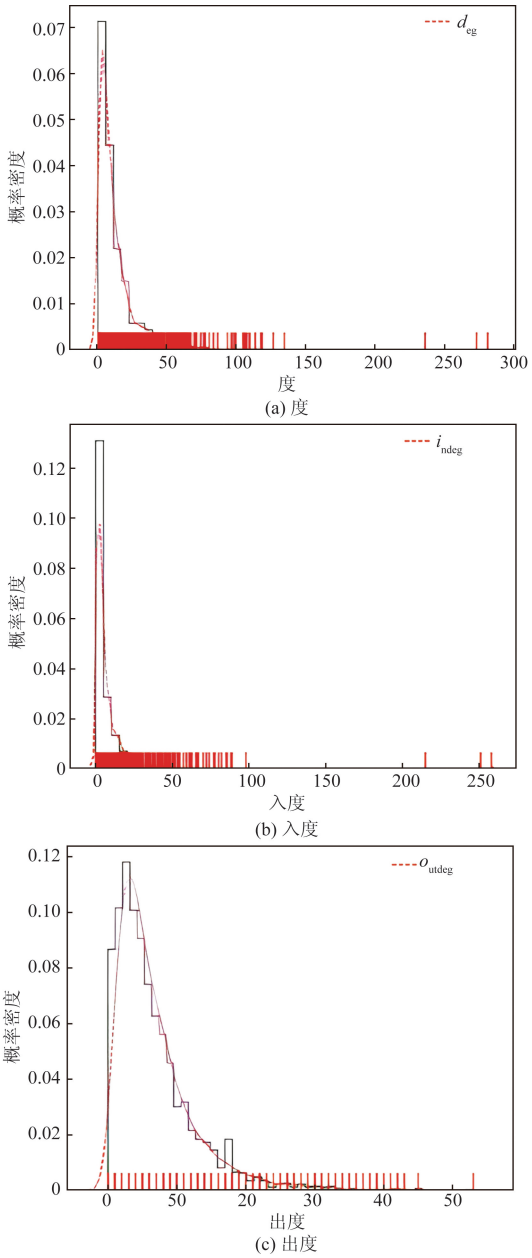


图 1　节点的度、入度及出度的分布情况

Fig. 1　Distribution of node's degree, in degree and out degree

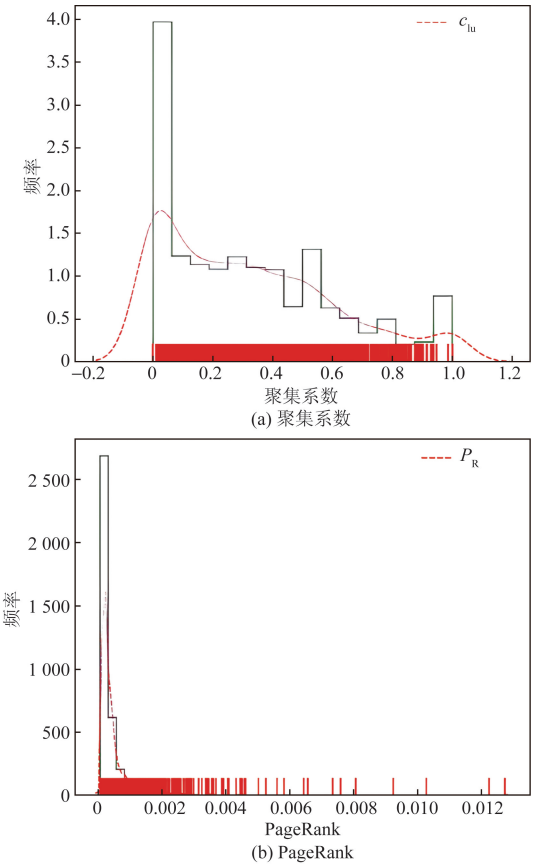


图 2　节点的聚类系数及 PageRank 的分布情况

Fig. 2　Node's distribution of clustering coefficients and PageRank

3.2　评价指标

选取准确性(Accuracy)、精确度(Precision)、召回率(Recall)及 F_1 分数(F_1-score)作为评估网页分类性能的评价指标。

3.3　实验对比方法和模型

对比的基准方法详细信息如下。

(1)传统方法[22]。使用 PageRank 基于网络链接结构进行网页分类。

(2)无属性图分类的基线方法[16]。通过节点的局部度特征和 SVM 进行分类。

(3)用 DeepWalk、Role2Vec 等图嵌入模型学习节点特征,再用逻辑回归进行分类[17]。

(4)BANE 模型结合 Weisfeiler-Lehman 邻近矩阵捕获节点链接和属性特征,用 SVM 进行分类预测[18]。

使用 5 种图嵌入模型来提取网页隐含关系特征,详细信息如下。

(1)DeepWalk[14]。算法分为两步,首先随机游走捕捉网络隐含信息,然后基于捕获的信息生成节点的低维特征向量。

(2)Node2Vec[15]。随机游走时同时考虑深度优先和广度优先邻域,以捕捉网络中的隐含信息。

(3)LINE[23]。通过学习网络中节点一阶和二阶近邻关系获取节点嵌入特征,以用于下游任务。

(4)SDNE[24]。使用深度自动编码器优化节点的一阶和二阶相似度,保留图的局部和全局结构,适用于稀疏网络。

(5)Struct2Vec[25]。不依赖近邻相似性,而是依据节点的空间结构相似性来学习网络中的隐含信息。

3.4 模型超参数配置

5 种图嵌入模型超参数设置如表 1 所示。分类模型参数配置如表 2 所示。

3.5 实验结果分析

3.5.1 利用隐含关系特征网页分类效果对比

本实验在真实网页链接数据集上进行,并采用 5 折交叉验证,实验结果如表 3 所示。其中最后5 行展示了本文方法的性能。以 DeepWalk* + XGBoost 为例,表示使用 DeepWalk 模型输出的 128 维节点 embedding 特征加上* 5 维统计结构特征作为 XGBoost模型的输入特征。

表 3 利用隐含关系特征网页分类的性能比较

Table 3 Performance comparison of web page classification using implicit relational features

方法名称	Accuracy	Precision	Recall	F_1-score
Pagerank + SVM[22]	0.570 4	0.589 6	0.621 1	0.579 9
LDP + SVM[16]	0.582 1	0.372 3	0.336 8	0.317 5
DeepWalk + LR[17]	0.696 5	0.685 2	0.609 2	0.630 1
BANE + SVM[18]	0.719 3	0.689 2	0.572 5	0.607 6
DeepWalk* + XGBoost	0.977 1	0.980 4	0.936 1	0.947 8
LINE* + XGBoost	0.977 1	0.982 2	0.922 0	0.946 0
Node2Vec* + XGBoost	0.985 4	0.915 3	0.924 7	0.919 5
SDNE* + XGBoost	0.968 8	0.797 8	0.810 2	0.803 6
Struct2Vec* + XGBoost	0.987 5	0.965 9	0.971 3	0.964 1

注:加粗文字表示性能最优的结果。

表 1 5 种图嵌入模型的超参数配置

Table 1 Hyperparameter configuration of graph embedding models

模型名称	模型参数	训练参数
DeepWalk	walk_length = 10, num_walks = 80, workers = 4	embed_size = 128, window_size = 5, workers = 3, iter = 3
Node2Vec	walk_length = 10, num_walks = 80, p = 0.25, q = 4, workers = 4, use_rejection_sampling = 0	embed_size = 128, window_size = 5, workers = 3, iter = 3
LINE	embedding_size = 128, order = 'second', negative_ratio = 5	batch_size = 1 024, epochs = 50, initial_epoch = 0, verbose = 2, times = 1
SDNE	hidden_size = [256, 128], alpha = 1×10^{-6}, beta = 5., nu1 = 1×10^{-5}, nu2 = 1×10^{-4}	batch_size = 3 000, epochs = 40, initial_epoch = 0, verbose = 2
Struct2Vec	walk_length = 10, num_walks = 80, workers = 4, verbose = 40, stay_prob = 0.3, opt1_reduce_len = True, opt2_reduce_sim_calc = True, opt3_num_layers = None, temp_path = './temp_Struct2Vec/', reuse = False	embed_size = 128, window_size = 5, workers = 3, iter = 5

注:walk length 为每次随机游走的步数;num walks 为每个节点的随机游走次数;workers 为并行进程数;embed size 为嵌入向量的维度;window size 为上下文窗口的大小;iter 为最大迭代次数。

表 2 分类模型的超参数配置

Table 2 Hyperparameter configuration of the prediction model

算法名称	算法参数配置
SVM	C = 1.0, kernel = 'rbf', degree = 3, gamma = 'scale', coef0 = 0.0, shrinking = True, probability = False, tol = 1×10^{-3}, cache_size = 200, max_iter = -1, decision_function_shape = 'ovr'
Logistic Regression	penalty = 'l2', C = 1.0, fit_intercept = True, intercept_scaling = 1, dual = False, tol = 1×10^{-4}, class_weight = None, solver = 'lbfgs', max_iter = 100, multi_class = 'auto', verbose = 0
RandomForest	n_estimators = 100, criterion = "gini", max_features = "auto", min_impurity_decrease = 0., bootstrap = True, verbose = 0, ccp_alpha = 0.0, min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0
GradientBoost	loss = 'deviance', learning_rate = 0.1, n_estimators = 100, subsample = 1.0, criterion = 'friedman_mse', min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0., max_depth = 3, min_impurity_decrease = 0., validation_fraction = 0.1, tol = 1×10^{-4}
XGBoost	max_depth = 3, learning_rate = 0.1, n_estimators = 100, verbosity = 1, silent = None, objective = "binary:logistic", booster = 'gbtree', n_jobs = 1, min_child_weight = 1, subsample = 1, colsample_bytree = 1, colsample_bylevel = 1, colsample_bynode = 1, reg_alpha = 0, reg_lambda = 1, scale_pos_weight = 1, base_score = 0.5

注:参数 C 为正则化参数,控制模型对误差的容忍度;kernel 为核函数类型;degree 为多项式核函数的次数;gamma 为核函数系数;coef0 为核函数中的独立项;shrinking 为是否启用 shrinking 启发式方法;probability 为是否启用概率估计;tol 为停止训练的容忍度;cache size 为内核缓存大小;max iter 为最大迭代次数,负值表示无限制;decision function shape 为多分类情况下决策函数的形状。

从表 3 可以发现,所有方法在 Accuracy、Precision、Recall 及 F_1-score 这 4 个评价指标上都取得了 0.5 以上的表现。这验证了利用隐含关系特征进行网页分类的有效性。同时,本文方法在网页分类的整体效果上展现出了更优的性能。其中,Struct2Vec* + XGBoost 方法在 Accuracy、Recall 及 F_1-score 这 3 个指标上均取得了最佳成绩,而 LINE* + XGBoost 方法则在 Precision 上表现最为突出。与其他基线方法相比,提出的方法在 4 个核心指标上均实现了显著的进步,这证明即使在缺乏网页显式特征的情况下,通过深入挖掘和利用节点在网页链接网络中的隐含关系特征信息,依然能够取得出色的网页分类效果。

3.5.2　网页分类中隐含关系特征贡献度分析

通过 XGBoost 训练 5 种图嵌入模型得到的特征矩阵,可以确定每个特征维度的重要性,如图 3 和图 4所示。分析这些指标有助于直观理解不同特征提取方法中前几位关键数据特征与网页分类结果的联系。图 3 以 DeepWalk 和 DeepWalk* 为例,展示了网页隐含关系特征的重要性分布。两者都显示

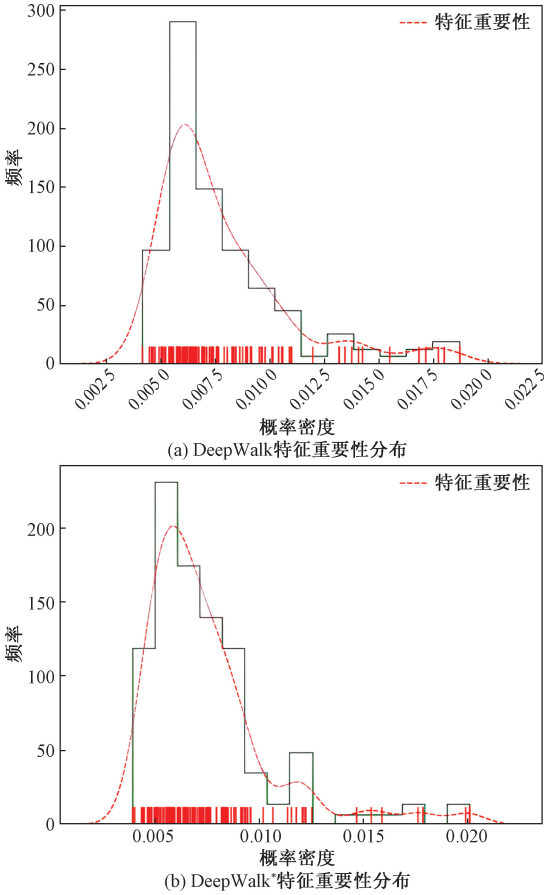


图 3　隐含关系特征重要性分布情况对比
(以 DeepWalk 和 DeepWalk* 为例)

Fig. 3　Comparison of feature importance distribution (taking DeepWalk and DeepWalk* as examples)

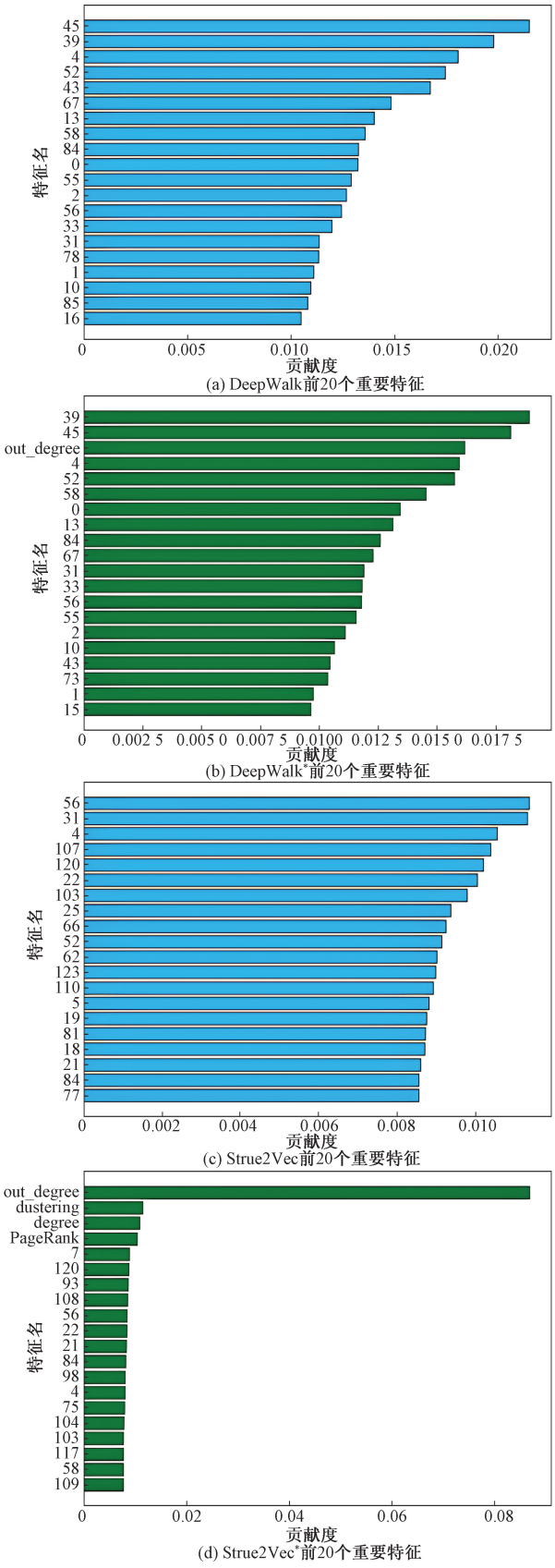


图 4　不同特征提取方案下的特征重要性对比
(DeepWalk 与 Struct2Vec)

Fig. 4　Comparison of feature importance under different feature extraction schemes (DeepWalk and Struct2Vec)

出均匀的特征重要性分布,呈现正偏厚尾型。这与3.1节讨论的原网页链接网络特征分布一致(图1、图2),这表明图嵌入模型将高维网页链接网络映射到低维空间时,保证了隐含关系特征分布的一致性。

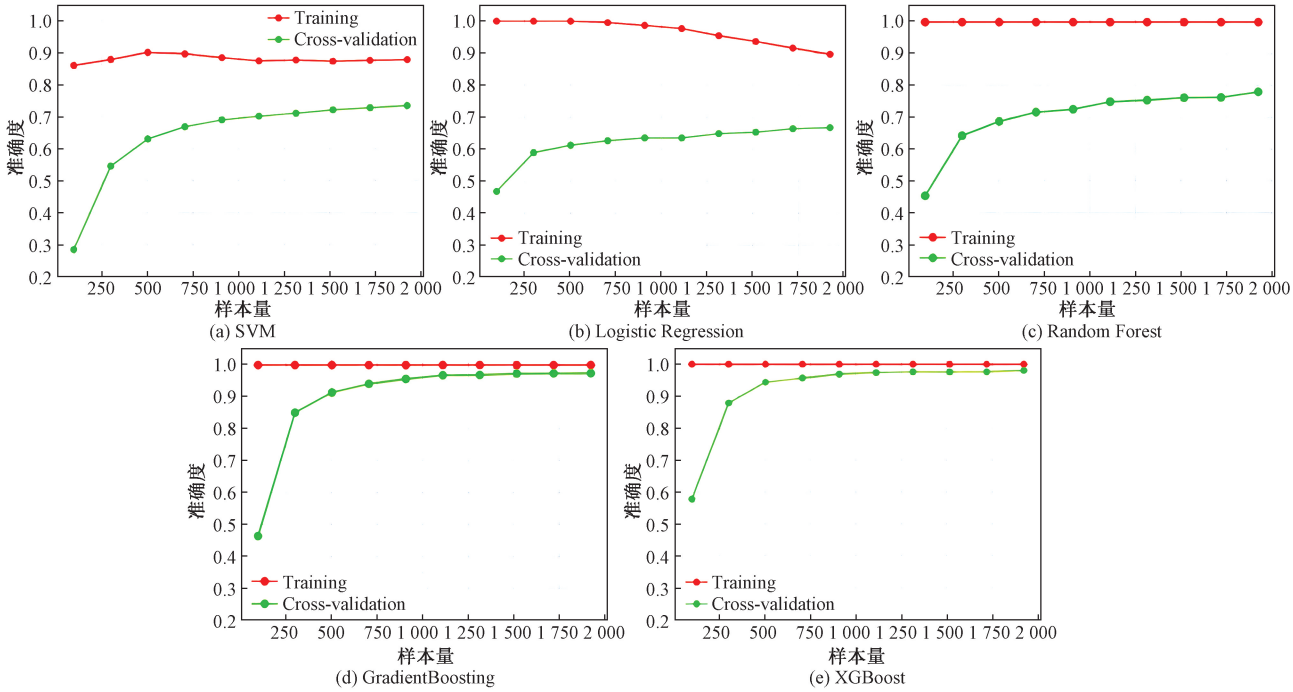
图4比较了DeepWalk和Struct2Vec两种特征提取方法的特征重要性。通过排序影响网页分类的特征,可以观察到关键特征与分类结果的关系。图4(a)和图4(b)对比了DeepWalk和DeepWalk*的前20个重要特征,而图4(c)和图4(d)对比了Struct2Vec及Struct2Vec*的前20个重要特征。结果显示,DeepWalk和Struct2Vec的前20个特征完全不同,反映了它们在训练数据抽取时的不同策略,导致两者之间特征分布的差异性,但两者都实现了将节点在图中高维隐含语义映射到低维空间的功能。同时,在缺乏网页显式内容特征的情况下,出度、聚集系数、度和PageRank等特征对分类结果影响显著,这些特征分别表示网页在链接网络中的位置、重要性和网页与其他相邻网页的链接情况。此外,相较于原始DeepWalk和Struct2Vec,DeepWalk*和Struct2Vec*通过结合节点连接性、重要性、中心性等统计特征,提供了更丰富的特征向量,增强了特征对网页分类结果的可解释性。

3.5.3 分类模型对网页分类效果的对比分析

本实验探讨了不同分类模型对网页分类效果的影响,同时保持节点特征提取方案一致,以增强实验结果的可比性。实验中,除文献[17-18]采用的传统LR和SVM分类模型,还引入XGBoost、GradientBoost等集成学习模型进行网页链接网络节点分类。图5展示了各模型学习曲线,表4和图6详细对比模型性能。

从图5可以看出,集成模型XGBoost和GradientBoost在分类效果上显著优于SVM、LR和RandomForest。具体来说,SVM、LR和RandomForest在样本量超过2 000时收敛效果不佳,训练集与交叉验证集性能差异显著。特别是LR,在样本量增加时训练集性能下降。而XGBoost和GradientBoost在样本量500~750时已展现良好拟合效果,且随样本量增加交叉验证性能稳定,未出现过拟合,对样本量依赖低。进一步对比XGBoost和GradientBoost,两者训练集性能均达100%,且XGBoost拟合速度更快,交叉验证性能更优。这表明在网页链接网络节点分类任务中,XGBoost比GradientBoost效率和准确性更高。

经过对表4和图6的细致分析,可以发现:在统一的节点特征提取方案下,不同分类模型的分类性能确实存在显著差异。从图6(a)可以看出,在此网页链接网络数据集上,XGBoost模型表现最佳,其次是GradientBoost模型,然后是RandomForest、SVM和LR。具体来说,在DeepWalk*、LINE*、Node2Vec和SDNE 4种特征提取方案中,LR的性能指标普遍



Training为训练集;Cross-validation为交叉验证集

图5 各分类模型的学习曲线对比(Struct2Vec*特征提取方案下)

Fig. 5 Comparison of learning curves for each classification model (under Struct2Vec* feature extraction scheme)

表 4　不同分类模型的分类性能对比

Table 4　Comparison of classification performance of different classification models

节点特征提取方法	分类方法	Accuracy	Precision	Recall	F_1-score
DeepWalk*	LR	0.696 5	0.685 2	0.609 2	0.630 1
	SVM	0.719 3	0.689 2	0.572 5	0.607 6
	RandomForest	0.779 6	0.694 5	0.547 0	0.588 1
	GradientBoost	0.962 6	0.802 6	0.803 9	0.802 7
	XGBoost	0.977 1	0.922 4	0.887 6	0.894 8
LINE*	LR	0.634 1	0.558 7	0.525 0	0.534 3
	SVM	0.700 6	0.665 4	0.516 8	0.549 1
	RandomForest	0.742 2	0.620 1	0.530 2	0.552 8
	GradientBoost	0.983 4	0.859 1	0.870 2	0.862 5
	XGBoost	0.970 9	0.940 3	0.913 4	0.920 9
Node2Vec*	LR	0.663 2	0.554 2	0.517 1	0.526 2
	SVM	0.719 3	0.663 1	0.540 7	0.571 9
	RandomForest	0.756 8	0.732 7	0.598 0	0.638 8
	GradientBoost	0.962 6	0.838 2	0.815 7	0.824 1
	XGBoost	0.973 0	0.916 7	0.880 2	0.882 4
SDNE*	LR	0.667 4	0.611 6	0.548 3	0.569 0
	SVM	0.638 3	0.620 7	0.455 3	0.489 8
	RandomForest	0.754 8	0.647 5	0.531 4	0.551 7
	GradientBoost	0.962 6	0.802 3	0.798 8	0.800 2
	XGBoost	**0.983 4**	0.901 1	0.923 0	0.906 7
Struct2Vec*	LR	0.303 5	0.202 0	0.208 1	0.201 0
	SVM	0.345 1	0.281 7	0.204 3	0.195 4
	RandomForest	0.596 7	0.391 7	0.344 1	0.345 5
	GradientBoost	0.973 0	0.894 2	0.918 2	0.899 9
	XGBoost	0.979 2	**0.985 2**	**0.939 1**	**0.951 1**

注:加粗文字表示性能最优的结果。

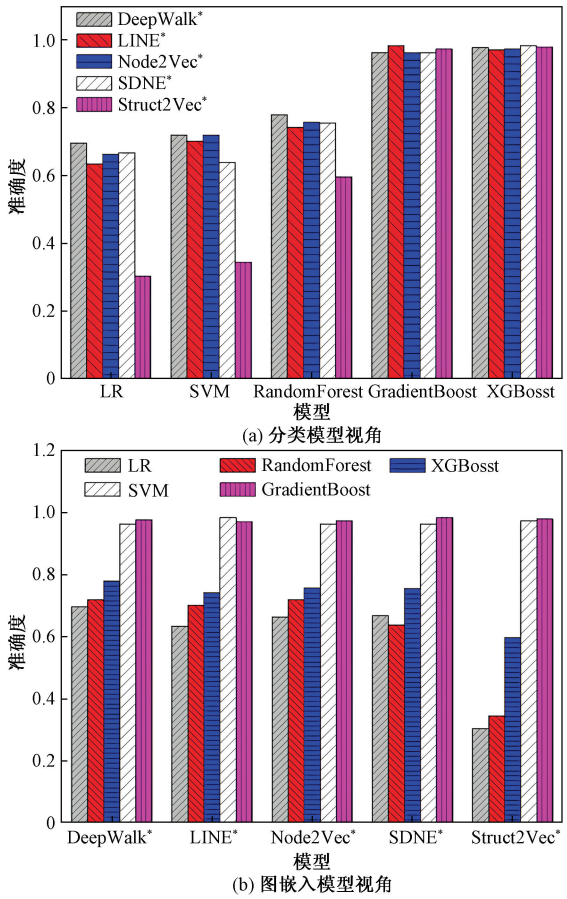


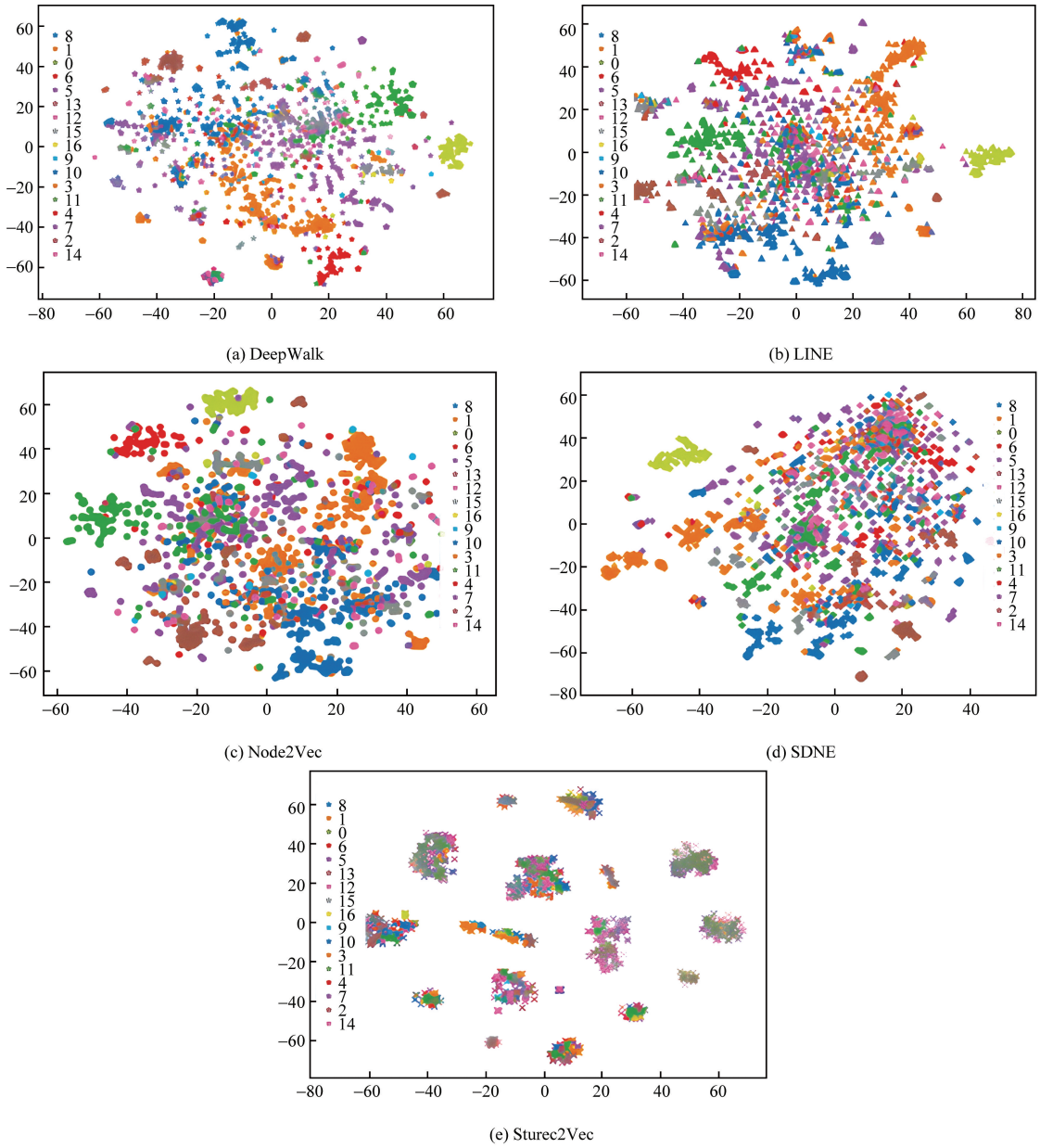
图 6　各分类模型的分类性能对比

Fig. 6　Comparison of the classification performance of each classification model

在[0.5 +,0.6 +]区间,而 SVM 和 RandomForest 模型性能稍高,处于[0.5 +,0.7 +]区间。相比之下,GradientBoost 和 XGBoost 这两个基于 boosting 的集成模型性能显著优于其他模型,大部分指标位于[0.8 +,0.9 +]区间,显示出对传统模型的明显优势。然而,对比表 4 的数据,注意到一个异常现象:如图 6(b)所示,LR 和 SVM 模型在 DeepWalk*、LINE*、Node2Vec 和 SDNE 4 种特征提取方案上的性能约为 0.6 +,但当应用到 Struct2Vec* 特征提取方案时,性能骤降至 0.2 +。Struct2Vec + XGBoost 的组合则取得 0.93 + 的优异性能。这一结果促使进一步探究其背后的原因。

通过将 t-SNE 降维技术[26-27]应用于 DeepWalk、LINE、Node2Vec、SDNE 和 Struct2Vec 共 5 种图嵌入方法提取的节点特征,得到了二维空间的散点图,如图 7 所示。图 7(a) ~ 图 7(d)显示,属于不同类别(共 17 类)的节点特征呈现出明显的聚集性,这表明即使经过降维处理,DeepWalk、LINE、Node2Vec 和 SDNE 提取的节点特征在二维空间中仍然保持着较为明显的空间分类特征。与此相反,图 7(e)显示 Struct2Vec 提取的节点特征在二维空间上分布并不呈现出明显的聚集规律,不同类别的节点特征往往无规律地聚集在一起,缺乏明显的分类界限。这一现象解释了 Logistic Regression 和 SVM 在 Struct2Vec* 特征上的性能会大幅下降的原因在于这些模型更适合处理具有明显线性或空间分布规律的特征,当面对 Struct2Vec* 这类提取的节点特征不具有明显线性或空间分布规律时,这些模型的预测效果会受到较大影响,导致性能骤降。

DeepWalk、LINE、Node2Vec 和 SDNE 4 种方法在原理上相似,主要区别在于训练样本的抽取方式,且在模型训练阶段多采用 Skip-gram 或 CBOW 结构,导致它们提取的特征在二维空间中的分布特征差异不大。因此,这些特征与 SVM 及 Logistic Regression 结合时,模型性能差异不大(表 4)。面对 Struct2Vec 特征,SVM 和 Logistic Regression 的分类效果显著下降,而 XGBoost 取得 0.93 + 的优异性能。是因为 XGBoost 通过迭代构建多个弱基模型,并利用加法模型与前向分步算法优化学习过程,使



图例处的数字 0 ~ 16 为节点(网页)类别编号, 共 17 个类别

图 7 节点特征降维后在二维空间上的散点分布情况

Fig. 7 Scatter distribution of node features on two-dimensional space after dimensionality reduction

其能更好地适应非线性或非空间分布规律的特征,因此在 Struct2Vec 特征上表现出色。

实验结果表明,在真实网页链接网络节点分类任务中,集成模型尤其是 XGBoost 展现出卓越的分类性能。同时,实验也提示,在相同的节点关系特征背景下,通过大量实验尝试不同的特征提取和分类模型组合,是提升网页链接网络节点分类性能的关键。

4 结论

(1)针对缺乏显式内容特征背景下的网页多分类问题进行了研究。通过 graph embedding 模型抽取节点(网页)在网页链接网络中的隐含关系特征,并与节点的集聚系数、PageRank 值等统计学结构特征拼接,共同构成节点的稠密特征向量,达到保留节点隐含关系特征同时将高维网页链接数据转化为低维特征向量的目的;其次,在构建节点组合特征的基础上,利用 XGBoost 等集成学习模型对节点进行分类预测,并分析了模型拟合效果及泛化能力,验证了适配合适分类模型的必要性。

(2)在真实的网页链接网络数据集上实验结果表明:在缺乏网页显式内容特征情况下,利用节点在网页链接网络中的隐含关系特征信息,取得了较好的网页分类效果,在准确率、精准率、查全率及 F_1

值4项指标上均优于已有方法,同时模型拟合收敛速度较快,也能够很好的适应样本量较小的应用场景。研究成果探索了网页链接网络中的隐含关系特征对网页分类的影响,尤其为无法获取到网页显式内容特征情况下的网页分类提供了新的视角。在隐私保护日益受到重视的今天,本文方法能够在不依赖于个人隐私信息的情况下进行网页分类,可以更准确地识别过滤网页,对推动中国在数据挖掘、网络安全、信息检索等领域的发展具有重要意义。

(3)由于数据源获取方式及模型离线训练模式的限制,导致模型迭代更新的计算成本较高,所以模型部署后需要更多的关注特征、标签漂移对模型性能的影响。在未来的研究中,计划尝试将模型在不同类型的网页链接数据集上进行测试,以验证模型的泛化性。

参 考 文 献

[1] 王法玉, 于晓文, 陈洪涛. 基于欠采样和多层集成学习的恶意网页识别[J]. 计算机工程与设计, 2024, 45(3): 669-675.
Wang Fayu, Yu Xiaowen, Chen Hongtao. Malicious web page recognition based on undersampling and multi-layer ensemble learning [J]Computer Engineering and Design, 2024, 45(3): 669-675.

[2] 张明杰, 肖奇荣, 朱烨行. 基于XGBoost模型的融合多特征微博信息传播预测方法[J]. 科学技术与工程, 2023, 23(10): 4279-4285.
Zhang Mingjie, Xiao Qirong, Zhu Yehang. Prediction method of microblog information dissemination based on XGBoost model and multi-feature fusion [J]. Science Technology and Engineering, 2023, 23(10): 4279-4285.

[3] 翁彬月, 秦永彬, 黄瑞章, 等. NEMTF: 基于多维度文本特征的新闻网页信息提取方法[J]. 计算机应用研究, 2022, 39(4): 1043-1048.
Weng Binyue, Qin Yongbin, Huang Ruizhang, et al. NEMTF: method of news Web content extraction based on multi-dimensional text features[J]. Application Research of Computers, 2022, 39(4): 1043-1048.

[4] 周文文, 韩斌, 黄树成. 结合文本语义图和词频统计的网页分类算法研究[J]. 计算机与数字工程, 2020, 48(6): 1265-1268, 1313.
Zhou Wenwen, Han Bin, Huang Shucheng. Research on web page classification algorithm combining text semantic graph and word frequency statistics[J]. Computer and Digital Engineering, 2020, 48(6): 1265-1268, 1313.

[5] 耿宜鹏, 鞠时光, 蔡文鹏, 等. 基于Skip-PTM的网页主题分类与主题变迁的研究[J]. 小型微型计算机系统, 2020, 41(7): 1395-1399.
Geng Yipeng, Ju Shiguang, Cai Wenpeng, et al. Research on topic classification and topic change of web pages based on Skip-PTM [J]. Journal of Chinese Computer Systems, 2020, 41(7): 1395-1399.

[6] 冯健, 张莹. 基于文档对象模型结构聚类的钓鱼网页检测方法[J]. 科学技术与工程, 2018, 18(23): 81-89.
Feng Jian, Zhang Ying. A detection method for phishing webpage based on DOM structure clustering[J]. Science Technology and Engineering, 2018, 18(23): 81-89.

[7] Deng L, Du X, Shen J Z. Web page classification based on heterogeneous features and a combination of multiple classifiers[J]. Frontiers of Information Technology & Electronic Engineering, 2020, 7: 995-1004.

[8] 淮晓永, 韩晓东, 高若辰, 等. 一种自适应网页结构化信息提取方法[J]. 电子技术应用, 2020, 46(12): 97-102.
Huai Xiaoyong, Han Xiaodong, Gao Ruochen, et al. An adaptive web page structured information extraction method[J]. Application of Electronic Technique, 2020, 46(12): 97-102.

[9] 洪良怡, 朱松林, 王轶骏, 等. 基于卷积神经网络的暗网网页分类研究[J]. 计算机应用与软件, 2023, 40(2): 320-325, 330.
Hong Liangyi, Zhu Songlin, Wang Yijun, et al. Darknet web page classification based on convolutional neural network[J]. Computer Applications and Software, 2023, 40(2): 320-325, 330.

[10] 张紫妍, 韩斌, 姜元昊, 等. 融合差分进化的网页暗链集成分类检测方法[J]. 计算机仿真, 2024, 41(4): 391-396.
Zhang Ziyan, Han Bin, Jiang Yuanhao, et al. Integrated Classification and detection method of web page hidden hyperlink based on differential evolution [J]. Computer Simulation, 2024, 41(4): 391-396.

[11] 杨胜杰, 陈朝阳, 徐逸, 等. 基于深度学习与特征融合的恶意网页识别方法研究[J]. 信息安全学报, 2024, 9(3): 176-190.
Yang Shengjie, Chen Zhaoyang, Xu Yi, et al. Research on malicious web page identification method based on deep learning and feature fusion [J]. Journal of Cyber Security, 2024, 9(3): 176-190.

[12] Giamphy E, Guillaume J L, Doucet A, et al. A survey on bipartite graphs embedding[J]. Social Network Analysis and Mining, 2023, 13(1): 54.

[13] 李青, 王一晨, 杜承烈. 图表示学习方法研究综述[J]. 计算机应用研究, 2023, 40(6): 1601-1613.
Li Qing, Wang Yichen, Du Chenglie. Survey on graph representation learning methods[J]. Application Research of Computers, 2023, 40(6): 1601-1613.

[14] Perozzi B, AL-Rfou R, Skiena S. DeepWalk: online learning of social representations[C]//Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014: 701-710.

[15] Grover A, Leskovec J. Node2Vec: scalable feature learning for networks[C]//Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 855-864.

[16] Cai C, Wang Y. A simple yet effective baseline for non-attributed graph classification[C]//Proceedings of the International Conference on Learning Representation. New York: ACM, 2018: 701-710.

[17] Rozemberczki B, Kiss O, Sarkar R. Karate club: an api oriented open-source python framework for unsupervised learning on graphs [C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM, 2020:

3125-3132.

[18] Yang H, Pan S, Zhang P, et al. Binarized attributed network embedding[C]//Proceedings of the International Conference on Data Mining. Piscataway, NJ: IEEE, 2018: 1476-1481.

[19] 张子威, 王鑫, 朱文武. 图神经架构搜索综述[J]. 计算机学报, 2023, 46(7): 1532-1552.
Zhang Ziwei, Wang Xin, Zhu Wenwu. Graph neural architecture search: a survey[J]. Chinese Journal of Computers. 2023, 46(7): 1532-1552.

[20] Chen T, Guestrin C. XGBoost: a scalable tree boosting system [C]//Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 785-794.

[21] 孙辰星, 刘伟, 卢彬, 等. 多视角网页分类数据集构建及性能评估[J]. 南京大学学报(自然科学), 2024, 60(3): 406-415.
Sun Chenxing, Liu Wei, Lu Bin, et al. Multi-View webpage classification dataset construction and evaluation[J]. Journal of Nanjing University(Natural Science), 2024, 60(3): 406-415.

[22] Pedroche F. A Model to Classify users of social networks based on PageRank[J]. International Journal of Bifurcation and Chaos, 2012, 22(7): 93-106.

[23] Tang J, Qu M, Wang M, et al. LINE: large-scale information network embedding[C]//Proceedings of the 24th International Conference on World Wide Web. New York: ACM, 2015: 1067-1077.

[24] Wang D, Cui P, Zhu W. Structural Deep network embedding [C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1225-1234.

[25] Ribeiro L F R, Saverese P H P, Figueiredo D R. Struct2Vec: learning node representations from structural identity[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2017: 385-394.

[26] Ruitmark V D, Billeter M, Eisemann E. An efficient dual-hierarchy t-SNE minimization[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(1): 614-622.

[27] 谢斌, 徐燕, 王冠超, 等. t-SNE 最大化的自适应彩色图像灰度化方法[J]. 中国图象图形学报, 2024, 29(8): 2333-2349.
Xie Bin, Xu Yan, Wang Guanchao, et al. Adaptive decolorization method based on t-SNE maximization[J]. Journal of Image and Graphics, 2024, 29(8): 2333-2349.