基于特征的社交网络恶意用户检测

施雯青 [1]
SHI Wenqing

摘 要 近年来，恶意用户的检测问题已成为社交网络安全领域的研究热点。当前的恶意用户检测方法多依赖基于内容特征的构建，但随着用户隐私保护需求的日益增加以及恶意用户行为的隐蔽性，这种依赖单一特征的检测方法面临诸多挑战。为此，文章从多个角度利用节点交互行为、节点拓扑分布以及邻居聚合来进行特征构建；并且，创新性地提出了一个边缘检测模型，将边作为整体进行特征构建，通过利用节点之间边缘的特征来识别节点标签，有效提高了恶意用户检测的准确性。并基于真实数据集构建了多个分类器模型并进行评估，验证了不同特征构建方法的有效性。

关键词 社交网络；拓扑分布；邻居聚合；边缘检测；恶意用户检测

doi：10.3969/j.issn.1672-9528.2025.02.027

0 引言

近年来，Twitter、Facebook等在线社交网络规模持续扩大，社交媒体平台用户数量不断增长，其为用户提供跨时空交流渠道，是日常生活重要部分。但社交网络的便利性也让恶意用户（如垃圾邮件发送者）有了滋生土壤，他们常创建虚假账户散布广告、发布含不良内容链接或向合法用户发恶意信息以窃取个人隐私数据[1]。因此，有效检测和应对这些垃圾邮件发送者成社交网络安全研究的关键课题。

恶意用户检测问题通常被建模为一个二元分类问题[2]。特征构建方法广泛应用于此领域，例如 Zafarani 等人[3]通过分析注册名的复杂度和性别等属性特征，成功检测 Twitter 网络中的恶意用户。Perez 等人[4]挖掘内容语义特征，发现恶意用户更偏爱使用肯定语气，且会使用较多关于时间上的词汇。Kumar 等人[5]利用发布消息频次、时间间隔等特征识别 Twitter 中垃圾信息制造者或通信网络中异常用户。Fire 等人[6]利用度、聚类系数、社区存在数、平均度特征以及偏离度特征检测网络中的虚假用户。但随着科技的发展，恶意用户伪装账号资料信息，造成虚假信息辨识越来越困难，恶意用户行为越来越隐蔽[7]，因此需要从多个角度结合不同特征进行考虑。

本文从多角构建用户特征以识别恶意用户、评估特征表现。主要贡献：一是为节点构建综合特征集，含行为、拓扑特征，捕捉多维信息；二是聚合邻居特征，强化节点表征、提升检测效果；三是提出新边缘检测模型，整体构建边特征，借边缘特征识别边、节点标签。

1. 南京审计大学计算机学院 江苏南京 210000

1 基于特征的社交网络恶意用户检测

本文从多个角度构建用户特征，以实现对恶意用户的有效识别，并评估不同特征在用户检测任务中的表现。具体而言，本文的主要贡献包括以下几个方面：

（1）为每个节点构建了综合的特征集，涵盖了节点的行为特征和拓扑特征，充分捕捉节点的多维信息。

（2）通过聚合节点邻居的特征，进一步增强了对节点的表征能力，提升了模型的检测效果。

（3）提出一种新的边缘检测模型，将边（连接）作为整体进行特征构建，利用节点之间边缘特征来识别边的标签，从而有效检测节点的标签。

1.1 特征构建

1.1.1 用户行为特征

关注数、粉丝数：在社交网络中，正常用户的关注数和粉丝数通常较为均衡，反映了其社交互动的特性。将节点的粉丝数记为 $N_{\text{followers}}$，关注数可记为 $N_{\text{followerings}}$。

关注者比例（follower ratio，FR）：相比于正常用户，恶意用户通过频繁地关注其他账号，来获得用户关注，从而让粉丝数增长，伪装身份。构建关于关注数和粉丝数的比例的特征：

$$\text{FR}=\frac{N_{\text{followers}}}{N_{\text{followers}}+N_{\text{followerings}}} \tag{1}$$

推文数：正常用户的推文内容通常具有一定价值，反映其自然的互动和信息分享行为。该特征可以表示为 $\text{num}_{\text{n\_tweet}}$。

推文转发比例（tweet retweet ratio，TRR）：恶意用户往

往不愿花费精力编写原创内容，更倾向于通过转发其他推文来保持活跃。本文用 $\text{num}_{\text{retweet}}$ 表示推文中转发的推文数量：

$$\text{TRR}=\frac{\text{num}_{\text{retweet}}}{\text{num}_{\text{n\_tweet}}} \tag{2}$$

正常用户和恶意用户在发送推文时常常会在时间上不呈现不同的方式。定义用户共发送了 $|T|$ 次推文，$T=\langle t_1,t_2,\ldots t_n\rangle$ 表示为节点发送推文的时间序列，用户的一阶时间序列可表示为$\Delta T=\langle \Delta t_1 \Delta t_2,\ldots \Delta t_{n-1}\rangle$，$\Delta t_r=t_{r+1}-t_r$，其中$1\le r\le n-1$。

最短时间间隔（minimum time interval，MTI）：时间间隔指的是同一用户相邻两条推文发送之间的时间差。短时间内的频繁推文可能是用户的恶意行为。

$$\text{MTI}=\min(\langle \Delta t_1,\Delta t_2,\ldots \Delta t_{n-1}\rangle) \tag{3}$$

时间差分序列（time-split sequence，TSS）：恶意用户可能会在短时间内发送多条推文，取$\Delta t_r\le\delta$，δ 表示为时间差的阈值，得到集合$N=\{\Delta t_r|\Delta t_r\le\delta\}$，则时间差分序列特征可以表示为：

$$\text{TSS}=\frac{|\{\Delta t_r|\Delta t_r\le\delta\}|}{n-1} \tag{4}$$

1.1.2 用户拓扑特征

目前已有大量的研究关注于分析网络结构在恶意用户检测方面的作用，网络的拓扑结构对于恶意用户的检测有非常大的帮助。

出度（out-degree，OutD）、入度（in-degree，InD）：社交网络中节点的出入度反映了节点在网络中的活跃程。

双向链接比例（bidirectional link ratio，BLR）：该特征意在反映 u 的所有邻居用户中，与该用户有往返交互的邻居的比例。使用 $N_{\text{in}}(u)$ 和 $N_{\text{out}}(u)$ 分别表示用户节点 u 的入度邻居集合和出度邻居集合，定义 u 的双向链接比例特征：

$$\text{BLR}=\frac{|N_{\text{in}}(u_i)\cap N_{\text{out}}(u_i)|}{|N_{\text{in}}(u_i)\cup N_{\text{out}}(u_i)|} \tag{5}$$

邻居出度均值（neighbor outdegree mean，NOM）：用来衡量节点邻居的活跃度，该特征可以定义为：

$$\text{NOM}=\frac{1}{|N_{u_i}|}\sum\frac{d_{u_j}^{\text{out}}}{d_{u_j}^{\text{out}}+d_{u_j}^{\text{in}}} \tag{6}$$

式中：N_{u_i}表示节点的所有邻居用户集合；$u_j\in N_{u_i}$，d_{u_j} 表示节点的度。

k-core[8]：节点中心性显节点在网络结构的重要性与影响力，经 k 核分解得用户节点核心度，代表所属 k-core 子图。

聚类系数[9]：量化节点连通、内聚性，正常用户系数高，邻居紧密；恶意用户系数低，多单向连接，邻居联系少。

1.1.3 用户聚合邻居特征

节点与邻居节点信息至关重要，传统方法多只看节点自身特征，没充分利用网络潜在信息。为此，提出基于节点对的邻居特征分布法，针对属性图，计算节点的 k-hop 邻居特征分布，既考虑节点自身属性，又结合邻居影响，使节点特征表达信息更丰富。对于每个节点 u，其邻居结合表示为 $N(u)$，该集合包含距离节点 u 一步的所有节点，节点 u 的 k 阶增强特征表示可以定义为：

$$X_u^k=f\left(X_v^{k-1}\,|\,v\in N(u)\right) \tag{7}$$

式中：x_v 为邻居节点 u 的特征，使用的聚合函数 $f(\cdot)$ 为均值聚合：

$$f(\cdot)=\frac{1}{|N(u)|}\sum_{v\in N(u)}X_v \tag{8}$$

接着，将节点的 k 阶特征表示与其原始特征拼接在一起：

$$X_u^{\text{final}}=X_u\oplus X_u^k \tag{9}$$

通过引入邻居的特征分布，可以增加特征信息的密度，使得节点的特征表示更加丰富。由于本文的数据集，节点的邻居节点非常密集，因此只考虑 2-hop 邻居的信息。

1.1.4 用户边缘特征

除构建节点特征检测外，还可考虑节点相连的边，因其承载交互信息，构建边特征、用边缘检测模型能增强检测效果，具体模型见第 2 节。

共同好友（common friends，CF）：两节点共同好友数量越多，相似度可能越高。

$$\text{CF}(u,v)=|N(v)\cap N(u)| \tag{10}$$

枢纽减小指数（hub depressed index，HDI)：衡量节点之间的相对相似程度，防止因某个节点度数过高导致相似程度偏高。

$$\text{HDI}(u,v)=\frac{|N(v)\cap N(u)|}{\max(d(v),d(u))} \tag{11}$$

枢纽增大指数（hub promoted index，HPI）：衡量节点之间的相对相似程度，更加注重低出度节点的作用。

$$\text{HPI}(u,v)=\frac{|N(v)\cap N(u)|}{\min(d(v),d(u))} \tag{12}$$

1.2 边缘检测模型

在社交网络里，正常用户倾向与同类交互，其发出信息的接收者多为其认可的正常用户，所以正常用户发出的连接大概率是同质性边。基于此假设，将边视为整体，已知一端为正常用户时，可构建边的特征推断其属性。具体从已识别的正常用户出发，以之为种子节点构建社交网络图与邻居关系网。设社交网络图 $G=\langle V,E\rangle$，E 是边 $e=(v,u)$ 的集合，V_{known} 表示已知标签的正常用户节点集合，V_{unknown} 表示未知标签的节点集合。在已知一端为正常用户节点的前提下，检测另一端节点的标签问题，可以将该问题转化为通过构

建特征 $X_e=\{x_1,x_2,\cdots,x_n\}$，$n$ 表示特征维度，计算边的同质性概率分布，从而确定另一端节点的标签。因此本模型的检测目的为：在给定训练集边的情况下，通过建模两端节点对边的影响，计算边的同质性概率分布，为边 $e=(v,u)$ 分配一个二进制标签 y_e，$y_e\in\{0,1\}$，表示异质性边或同质性边，从而判断节点的标签：$p(y_e\mid X_e)\Rightarrow p(y_u\mid y_v,y_e)$，具体的流程如图 1 所示。

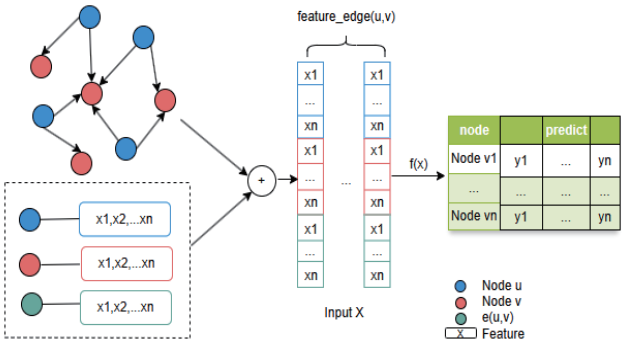

图 1 边缘检测模型

一个节点可能会与多个已知标签的节点相连，由于每个标签得出的过程都是独立的，因此选用多数投票法来决定最终的标签。模型得到节点的多个预测标签 $y_{u,v}$，$v\in N_u$，$y_{u,v}\in\{0,1\}$，选择出现次数最多的标签作为最终标签：

$$y_{u,\text{final}}=\arg\max_{y_u\in\{0,1\}}\left(\text{Count}\left(y_u\right)\right) \tag{13}$$

式中：Count(y_u) 为分别统计两个预测标签的次数。

2 实验与分析

2.1 实验数据集

用公共数据集 Twitter 1KS -10KN[10] 评估构建特征的有效性，它从 Twitter 中收集而来，含标记的恶意、正常用户，以及推文与用户交互次数。此数据集含有 1000 个恶意用户、10 000 个正常用户，推文 1 354 616 条，交互 2 287 930 次。

2.2 实验方法

为比较各特征在用户识别中的能力，全面评估了构建的特征。用户行为、拓扑、邻居聚合特征基于单个节点，用分类器评估；边缘特征基于边结构，用专门边缘检测模型实验。实验采用十折交叉验证，边缘检测模型新测试集由训练集种子节点与测试集节点间的边组成，导致节点不一致，因此对齐了节点特征与边缘检测模型的测试集。

2.3 实验评估指标

鉴于 1KS-10KN 数据集极度不平衡，用精确度召回曲线下面积（PRAUC）与 ROC 曲线下面积（ROCAUC）共同评价，数据不平衡时，前者评估更客观，后者可能存在欺骗性，最后用常用的 F_1 值、Recall 值作重要评价指标。

2.4 实验结果

2.4.1 实验一：基于节点特征的模型评估

通过构建的特征比较在不同分类器的检测结果。首先，针对用户行为特征和拓扑特征，分别进行独立检测，以便深入理解每种特征在恶意用户识别中的有效性，具体结果如表 1、表 2 所示。

表 1 用户行为特征

algorithm	Recall	F_1	ROCAUC	PRAUC
XGBoost	0.742 6	0.769 2	0.861 8	0.782 0
Bagging	0.687 9	0.746 1	0.837 3	0.763 7
决策树	0.626 4	0.621 3	0.795 6	0.636 7
K 近邻	0.699 5	0.746 9	0.840 5	0.764 8
逻辑回归	0.539 9	0.653 4	0.763 9	0.705 9

表 2 用户拓扑特征

algorithm	Recall	F_1	ROCAUC	PRAUC
XGBoost	0.765 0	0.783 1	0.848 0	0.805 8
Bagging	0.670 5	0.737 5	0.828 8	0.758 1
决策树	0.690 4	0.686 9	0.829 5	0.700 7
K 近邻	0.589 2	0.643 1	0.783 4	0.665 8
逻辑回归	0.635 2	0.660 1	0.813 0	0.685 1

接着，我们将这两类特征进行综合考虑，通过整合用户行为特征与拓扑特征，来评估准确性具体结果见表 3。

表 3 用户行为特征 + 用户拓扑特征

algorithm	Recall	F_1	ROCAUC	PRAUC
XGBoost	0.740 4	0.834 7	0.868 4	0.860 7
Bagging	0.784 6	0.840 7	0.888 3	0.854 5
决策树	0.727 3	0.737 1	0.852 2	0.748 8
K 近邻	0.564 9	0.662 9	0.775 0	0.704 3
逻辑回归	0.474 7	0.593 1	0.731 1	0.655 9

评估聚合邻居特征对恶意用户检测的贡献，进一步提升节点特征的表达能力和信息密度，具体结果如表 4 所示。

表 4 邻居聚合特征

algorithm	Recall	F_1	ROCAUC	PRAUC
XGBoost	0.831 7	0.872 7	0.912 1	0.882 6
Bagging	0.767 0	0.828 2	0.879 8	0.842 8
决策树	0.790 8	0.765 4	0.881 9	0.775 5
K 近邻	0.456 1	0.593 2	0.724 5	0.673 6
逻辑回归	0.655 0	0.704 3	0.817 3	0.724 0

实验表明，集成分类器在恶意用户检测上优于单一分类器。表 1、2 展示基于用户行为、网络拓扑特征的分类结果，单独用虽在部分场景有效，但因仅捕捉单一方面信息，识别受限。表 3 结合了用户行为特征与网络拓扑特征，显著提升了分类效果。表 4 进一步融合了用户自身和邻居节点的特征，丰富了特征表示的维度，实验表明引入邻居信息有效增强了模型的检测能力，让集成分类器性能大幅提升。

2.4.2 实验二：基于边缘特征的模型评估

接下来通过检测边缘特征来进行识别。我们通过检测边的标签来识别节点的标签。由于每个节点的入度存在显著差异，正常用户节点的入度通常远高于垃圾邮件节点的入度，这导致了数据的严重不平衡性。因此在选择邻居节点时，我们对节点的邻居数据集进行了采样，确保每个节点仅考虑其部分入度节点，具体的检测结果见表 5。

表 5 边缘特征 + 节点特征

algorithm	Recall	F_1	ROCAUC	PRAUC
XGBoost	0.830 4	0.882 0	0.922 9	0.892 1
Bagging	0.789 5	0.838 5	0.890 6	0.850 1
决策树	0.807 0	0.762 4	0.890 0	0.772 4
K 近邻	0.720 8	0.776 0	0.853 7	0.793 0
逻辑回归	0.783 6	0.644 2	0.863 7	0.752 0

从表 5 可知，构建的边缘检测模型性能优异，结合节点自身与所连边的特征，能全面捕捉行为、结构特征，在恶意用户检测中精度更高，表明考虑边缘信息更有效识别恶意用户，提升集合分类器检测效果。

3 总结

本文从多维度构建用户特征，含行为、拓扑、邻居聚合及边缘特征，针对边缘特征创新地借检测边属性反推节点标签检测恶意用户。在真实数据集用集成、单一分类器实验，结果显示 XGBoost 模型表现出色。未来需完善图结构用户特征探讨，且当前实验依赖监督学习，探索无监督学习对数据稀缺下恶意用户检测意义重大，能提供新的思路。

参考文献：

[1] LAZER D M J, BAUM M A, BENKLER Y, et al. The science of fake news[J]. Science, 2018,359(6380): 1094-1096.

[2] 唐冰聪 . 在线社交网络中恶意用户行为分析及检测方法 [D]. 南京 : 南京财经大学 , 2021.

[3] ZAFARANI R, LIU H. 10 Bits of surprise: detecting malicious users with minimum information[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. NewYork:ACM,2015: 423-431.

[4] PEREZ-ROSAS V, KLEINBERG B, LEFEVRE A, et al. Automatic detection of fake news[C]//Proceedings of the 27th International Conference on Computational Linguistics. Brussels: ACL, 2018:3391-3401.

[5] KUMAR S, CHENG J, LESKOVEC J, et al. An army of me: sockpuppets in online discussion communities[DB/OL].(2017-03-21)[2024-05-19].https://doi.org/10.48550/arXiv.1703.07355.

[6] FIRE M, KATZ G, ELOVICI Y. Strangers intrusion detection-detecting spammers and fake profiles in social networks based on topology anomalies[J]. Human journal, 2012(1): 9.

[7] 杨善林 , 王佳佳 , 代宝 , 等 . 在线社交网络用户行为研究现状与展望 [J]. 中国科学院院刊 , 2015, 30(2): 200-215.

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

[9] FAKHRAEI S, FOULDS J, SHASHANKA M, et al. Collective spammer detection in evolving multi-relational social networks[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.NewYork:ACM, 2015: 1769-1778.

[10] CHAO Y, HARKREADER R, ZHANG J L, et al. Analyzing spammer's social networks for fun and profit: a case study of cyber criminal ecosystem on twitter[C]//Proceedings of the 21st International Conference on World Wide Web. NewYork: ACM, 2012:71-80.

【作者简介】

施雯青（1999—），女，浙江湖州人，硕士研究生，研究方向：数据挖掘、大数据审计。

（收稿日期：2024-11-01）