一种改进图神经网络的虚假评论检测方法

袁紫烟,任勋益,黄家铭

(南京邮电大学 计算机学院、软件学院、网络空间安全学院,江苏 南京 210023)

摘 要: 电商平台中的虚假评论存在误导消费者购买决策、损害消费者和合法商家权益的问题。现有的虚假评论识别方法难以发现虚假评论之间的隐含关联。为了提高虚假评论检测的分类效果,提出一种基于TrustRank算法和图神经网络的虚假评论检测方法。首先,通过构建虚假评论相关特征,计算出特征的重要性分数;其次,结合改进TrustRank方法计算评论的可疑值,利用自适应邻域采样策略对图中节点进行有偏向和自适应地选择并生成目标节点的邻域,以此改进GraphSAGE的随机采样算法;最后,使用Yelp数据集对该模型进行验证。结果表明,TR-GraphSAGE模型相比于LSTM、TextCNN、GCN和GraphSAGE,在准确率、召回率与F1 3个方面分别提升了5.86%、15.01%和10.12%。TR-GraphSAGE模型能够降低噪音对预测的影响,保证邻域信息的质量和数量,从而提高关联特征表示质量,为虚假评论检测提供了一种新方法。

关键词: 虚假检测;TrustRank;图神经网络;特征工程;GraphSAGE

DOI: 10.11907/rjdk.231163 **开放科学(资源服务)标识码(OSID):**

中图分类号: TP391.1 **文献标识码:** A **文章编号:** 1672-7800(2024)003-0027-07

A Fake Reviews Detection Method Based on Improved Graph Neural Network

YUAN Ziyan, REN Xunyi, HUANG Jiaming

(*School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing* 210023, *China*)

Abstract: Fake reviews in e-commerce platforms mislead consumers' purchase decisions, and damage the rights and interests of consumers and legitimate businesses. The existing methods are difficult to find the implicit association between fake reviews. In order to improve the classification accuracy of fake reviews detection, a fake review detection method based on the TrustRank algorithm and graph neural network was proposed. Firstly, the features associated with the fake reviews were constructed, and the importance scores of these features were calculated. Secondly, to improve the random sampling algorithm of GraphSAGE, the suspicious values of fake reviews were calculated by the improved TrustRank method, which combined the adaptive neighborhood sampling strategy was used to select nodes in the graph with bias and adaptive and generate the neighborhood of target nodes. Finally, Yelp data set was used to verify the proposed model. The accuracy, recall and F1 of TR-GraphSAGE model were approximately 5.86%, 15.01%, and 10.12% better on average, respectively, than LSTM, TextCNN, GCN, and GraphSAGE. The TR-GraphSAGE model can eliminate the noise that affects the prediction, ensure the quality and quantity of the neighborhood, and thus improve the quality of the associated feature representation, which provides a new method for fake review detection.

Key Words: fake review detection; TrustRank; graph neural network; feature engineering; GraphSAGE

0 引言

随着互联网和电商平台的快速发展,商品评论已成为电商平台中不可或缺的一部分,极大地影响着消费者的购买决策[1-2],也被商家用于评估品牌感知效果和客户满意度水平[3]。然而,虚假评论在电商平台上逐渐增多,扰乱了市场秩序,出现劣币驱除良币的现象。例如,酒店评论评分降低1%,房间被预订的可能性会降低约2.6%[4];餐厅评分降低一分,销售收入就会降低5%~9%[5]。因此,如何

收稿日期: 2023-02-24

作者简介: 袁紫烟(1997-),女,南京邮电大学计算机学院、软件学院、网络空间安全学院硕士研究生,研究方向为虚假评论检测;任勋益(1973-),男,博士,南京邮电大学计算机学院、软件学院、网络空间安全学院副教授,研究方向为信息安全;黄家铭(1995-),男,南京邮电大学计算机学院、软件学院、网络空间安全学院硕士研究生,研究方向为虚假评论检测。本文通讯作者:袁紫烟。

在评论中筛选出虚假信息至关重要。

1 相关工作

虚假评论检测本质上是二元分类问题。现有方法主要分为3类:传统机器学习方法、深度学习方法和图神经网络方法。传统机器学习方法主要包括朴素贝叶斯算法、K-近邻算法和支持向量机等[6-8]。例如,Elmogy等[9]提出一种基于机器学习的虚假评论检测方法,该方法同时考虑了文本特征和行为特征,用于检测Yelp餐厅评论数据集中的虚假评论。王琢等[10]针对店铺类虚假评论特点,采用多种机器学习算法对店铺评论数据进行有监督的分类,进一步探索店铺类虚假评论的特征模式。Kontsewaya等[11]基于现成的带标签数据集,比较了朴素贝叶斯算法、k近邻算法、支持向量机、逻辑回归、决策树、随机森林等几种机器学习方法的分类性能,在基于语言特征的背景下,逻辑回归和朴素贝叶斯算法具有最好的表现。然而,机器学习方法利用虚假评论的内容特征和行为特征识别虚假评论,存在对高维特征处理能力差等局限性[12-13]。

因此,为了克服机器学习的局限性,学者开始采用深度学习、循环神经网络、卷积神经网络(Convolutional Neural Network,CNN)和长短期记忆网络(Long Short Term Memory Network,LSTM)等模型[14-17]。例如,Bathla等[18]提出一种结合CNN和LSTM的深度学习方法,相比于机器学习算法,该方法在Ott和Yelp数据集上的检测效果更好。陈宇峰[19]使用CNN-LSTM结合Doc2Vec和TF-IDF,在酒店评论数据集deceptive-opinion-spam-corpus的实验测试中正确率达到了93.1%。Qiu等[20]开发了一种深度匹配网络MIRD,在Yelp和Amazon数据集上的性能表现相比Transformer与LSTM等模型平均提高了8%。深度学习方法使用不同结构的深度神经网络进行端到端的学习,对虚假评论检测的准确率有较大提高,但其对文本的处理是序列化的,在发现虚假评论之间的隐含关联方面存在不足[21-24]。

近年出现的图神经网络(Graph Neural Networks,GNN)以其在分类精度上的优越性被应用于文本分类领域,具有提取结构表示能力强、鲁棒性较好、可有效聚合邻域特征等优点[25-26]。Song等[27]提出一种基于动态传播图的虚假评论检测方法,通过捕捉静态网络中缺失的动态传播信息,可有效提高虚假评论检测的准确率。Neisari等[28]通过自组织映射结合无监督学习聚类语义相似的单词,将文本转换为图像输入到CNN中进行训练,提高了在单域和多域上下文中的分类效果。曹东伟等[29]提出基于融合语义相似度的图卷积网络(Graph Convolutional Networks,GCN)虚假评论检测方法,在公开数据集上,该方法的准确率相对于CNN、LSTM及Text-GCN分别提升了7%、4.8%和1.3%。以上研究证明了图神经网络模型相比其他方法性能更出色,但其仍易受到邻域噪声的影响[30-32]。

因此,本文提出一种基于TrustRank算法的图神经网络虚假评论检测方法,通过对用户之间的关联信息进行建模,有效挖掘数据集的隐含信息。使用TrustRank改进图神经网络GraphSAGE的邻域采样策略,实现自适应邻域选择,从而改善现有方法在采样过程中忽视节点之间存在巨大差异、丢失包含更多信息的节点所带来的噪音干扰问题。TR-GraphSAGE模型在保证准确率的同时,可有效提高召回率。

2 改进图神经网络的虚假评论检测

本文提出的虚假评论检测框架如图1所示,包括数据预处理与特征提取,基于TrustRank的可疑值算法、TR-GraphSAGE算法进行图计算,以及模型评估对比。首先,对数据集进行预处理,并根据评论的文本内容与关联信息提取文本的行为特征和语言特征;接着,基于TrustRank提出可疑值算法,以计算评论图中节点的可疑值;之后,以节点的可疑值为依据改进GraphSAGE算法,提出TR-GraphSAGE算法;最后,使用精确率、召回率和F1指标对所提出的模型进行评估,并与LSTM、TextCNN、GCN和GraphSAGE模型进行比较。

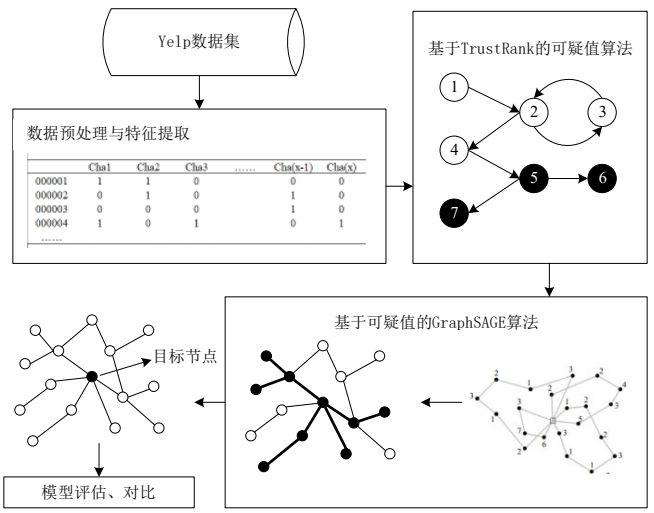

Fig. 1 Framework for fake review detection

图1 虚假评论检测框架

2.1 数据集与特征构建

2.1.1 数据预处理

使用的数据集为Mukherjee等[33]爬取Yelp收集的餐厅真实数据集,该数据集中评论者数量为16 930位,总评论数量为78 940条,其中真实评论69 468条,虚假评论9 472条。数据预处理对分类器预测结果会造成直接影响[34],对数据集中文本数据的预处理如下:①用单词number代替文本中的所有数字,用单词money代替所有$符号;②去除文本中的所有标点符号和特殊符号;③英语单词大写转小写;④进行词性还原,把单词还原为主要词形;⑤去

除停用词,其中包括在语言表达中无含义的语气词和助词,以及在语料库中出现频率很高,但特征并不明显,并且对语义提取无贡献的词,如“want”“out”等。

对数据集中其他特征的预处理如下:①用时间戳表示日期;②用整型表示原始数据中的店铺ID、评论ID和评论者ID;③丢弃没有行为信息的数据;④用“1”标识虚假评论,用“0”标识真实评论。

2.1.2　特征提取

目前,针对虚假评论的研究方向主要集中于文本特征和行为特征两个方面。其中,文本特征主要根据自然语言处理中的文本分类问题进行检测。基于行为特征的虚假评论研究需要从评论者的个人信息和行为中提取特征。在考虑使用的数据集和前人研究成果的基础上[35],分别从行为特征和文本特征两个角度分析数据并构建相关特征。

行为特征属于非语言特征中的一种,以下为基于行为构建的特征:

(1)创建天数(Creation Days,CD)。用户最后一次发布评论到注册账户日期之间的差值。量化公式如下:

$$CD = LastReview - JoinDate \tag{1}$$

其中,*LastReview*为用户最后一次发布评论的日期,*JoinDate*为用户注册账户的日期。

(2)总评论数量(Review Number,RN)。单个用户发布的总评论数量。

(3)平均发布比例(Average Publishing Ratio,APR)。用户总评论数量与用户活跃天数的比例,此特征揭示了评论者在活跃时间内平均发布的评论数量。量化公式如下:

$$APR(a) = N_r(a)/N(活跃时间) \tag{2}$$

其中,$N_r(a)$为用户的总评论数量,N(活跃时间)为总活跃天数。如果一个评论者在某天发布了至少一条评论,当天即为评论者活跃时间,活跃天数即为活跃时间之和。

(4)积极评论比例(Percentage of Positive Reviews,POS)。用户评分大于或等于4的评论占用户总评论数量的比例。量化公式如下:

$$POS(a) = N_r(reviewRating >= 4)/N_r(a) \tag{3}$$

(5)消极评论比例(Proportion of Negative Reviews,NEG)。用户评分小于或等于2的评论占用户总评论数量的比例。量化公式如下:

$$NEG(a) = N_r(reviewRating <= 2)/N_r(a) \tag{4}$$

(6)评分偏差(Rating Deviation,RD)。评论的评分与同一产品所有评论评分均值的差值。量化公式如下:

$$RD(a) = |r(a) - MEAN_r| \tag{5}$$

其中,$r(a)$表示单条评论评分,$MEAN_r$表示产品的评分均值。

(7)最大发布评论数量(Maximum Number of Posted Reviews,MPR)。用户单天内发布的最大评论数量。

(8)活跃时间窗口(Active Time Window,ATW)。此特征定义为用户第一次和最后一次评论的时间戳之差。量化公式如下:

$$ATW(a) = D_l(a) - D_f(a) \tag{6}$$

其中,$D_f(a)$为用户第一次发表评论的时间戳,$D_l(a)$为用户最后一次发表评论的时间戳。

(9)赞同数量(Approval Number,AN)。元数据中“usefulCount”、“coolCount”和“funnyCount”3个属性记录着用户发布的评论被其他用户点赞的次数。量化公式如下:

$$AN(a) = usefulCount(a) + coolCount(a) + funnyCount(a) \tag{7}$$

(10)好友数量(Number of Friends,FN)。发布此条评论的用户所拥有的好友数量。

文本特征又称为上下文特征,以下为基于语言文本构建的特征:①文本长度(Review Text Length,RL),评论文本的单词总数量;②文本平均自相似度(RCS),评论者自身发布所有评论相似度和的平均。量化公式如下:

$$RCS(a) = (\sum_i^n similarity(r_a, r_i))/n \tag{8}$$

其中,n为评论者发布评论的总数量,r_a为目标评论,r_i为除r_a外的其他评论。计算每条评论r_a与r_j的相似度,最后把比较后每条评论的相似度相加,再进行标准化处理。

2.2　TR-GraphSAGE模型构建

2.2.1　基于TrustRank的可疑值算法

将GraphSAGE算法直接用于虚假评论检测会存在噪音干扰和采样集合不稳定的问题。为此,提出基于TrustRank的可疑值算法,通过计算评论网络节点的可疑值来量化每个节点的可疑程度,针对可疑值高的邻域节点进行采样,去除影响虚假节点检测的可疑节点,从而保证邻域信息质量。

TrustRank的前身PageRank算法曾被谷歌用于搜索引擎来计算网页排名,其主要依赖于互联网页面间的链接关系[36]。页面所链接的质量越高,则排名越高。TrustRank相当于一个有偏向性的PageRank,即通过人工或某种规则选取种子集合,在信任指数传播过程中,更偏向于访问该集合及其周围页面。TrustRank算法通过定义初始的标签或种子,然后依靠web网络节点之间的相互连接来传播标签或种子中的信任指数。在迭代一定次数后或前一次和当前的信任指数几乎没变化时,网络达到稳定状态,终止算法执行。

参考TrustRank算法提出“可疑值”(Suspiciousness Value)的概念,使用符号SV来表示,用于量化节点为虚假节点的程度。

采用随机森林算法计算虚假评论中每个特征对初始可疑值的影响,并通过线性加权把所有特征组合起来,其分数作为初始可疑值。具体计算公式如公式(9)所示:

$$initSV(j) = \sum_i^n w_1 f_1 + w_2 f_2 + ... + w_n f_n \tag{9}$$

其中,$initSV(j)$为虚假节点j的初始可疑值,其考虑了虚假实体中各个特征的重要性对于初始可疑值计算的影响,w_i为第i个特征权重,f_i为第i个特征的具体数值。

接下来定义可疑值在评论网络中的传播公式(10)。

$$SV(r_i)=\alpha\sum_{r_j\in E(r_i)}\frac{SV(r_j)}{degree(r_j)}+(1-\alpha)\times\frac{initSV(i)}{N}\quad(10)$$

其中，α 为区间(0,1)的衰减因子，主要控制当前节点跳转到其他节点的概率；$E(r_i)$ 为与评论 r_i 用无向边连接的节点集合；$SV(r_j)$ 为与评论 r_i 连接的评论 r_j 的 SV 值；$|V|$ 为评论网络节点总数；$initSV(i)$ 为种子集合中第 i 个位置的可疑值，长度为 $|V|$；$degree(r_j)$ 为与评论 r_i 相连的评论 r_j 的度；N 为可疑种子的总数。在初始阶段，网络中每一个节点的可疑值都设置为对应节点的归一化初始可疑值 $\frac{initSV(i)}{N}$，用于迭代计算最终的可疑值。

算法 1：可疑值算法

输入：评论—评论矩 $\boldsymbol{T}$，迭代次数 n，种子数量 N，衰减因子 α，节点数量 $|V|$

输出：可疑值向量 $\boldsymbol{SV}$

1.$\boldsymbol{initSV}=0_{|V|}$

2.FOR j = 1 TO $|V|$ DO

3. $initSV(j)=\sum_i^n w_1f_1+w_2f_2+...+w_nf_n$

4.$\boldsymbol{initSV}=TopNNode(\boldsymbol{initSV})$

5.$\boldsymbol{initSV}=\frac{initSV(i)}{N}$，$\boldsymbol{SV}=\boldsymbol{initSV}$

6.FOR i = 1 TO n DO

7. $\boldsymbol{SV}=\alpha\cdot\boldsymbol{T}\cdot\boldsymbol{SV}_{i-1}+(1-\alpha)\cdot\boldsymbol{initSV}$

8.RETURN $\boldsymbol{SV}$

评论—评论网络的矩阵 $\boldsymbol{T}$ 是一个概率转移矩阵，E 为网络中所有边的集合。$T(i,j)$ 可以理解为节点 i 随机游走到节点 j 的概率，在矩阵 $\boldsymbol{T}$ 中表示为第 i 行第 j 列的元素。计算公式如下：

$$T(i,j)=\begin{cases}\dfrac{1}{degree(j)},(i,j)\in E\\ 0,\ 其他\end{cases}\quad(11)$$

首先输入评论—评论网络的矩阵 $\boldsymbol{T}$，迭代次数为 100 次，确保每个节点的可疑值基本不会发生变化。设置种子数量 N 为 100 个，衰减因子 α 为 0.85，节点数量为 $|V|$。可疑值算法流程如下：第一行将初始可疑值进行初始化；第二行对网络上的所有节点进行遍历；第三行将节点的特征权重和特征值进行加权求和得到初始可疑值；第四行对初始可疑值集合进行排序，并挑选出 100 个种子，再对种子的对应位置设置初始可疑值，其余位置设置为 0；第五行对得到的可疑种子向量进行归一化处理，使得初始可疑值落在(0,1)区间之间，并为网络中每一个节点赋予归一化初始可疑值；第六、七行基于 TrustRank 的改进算法以迭代的方式在节点之间传递可疑值，在每次迭代中，每个节点的可疑值从一个节点传递到其邻接点，种子节点的可疑值也随着评论网络不断传播，迭代结束后可以得到每个节点的最终可疑值。

2.2.2 TR-GraphSAGE 算法

虚假评论者与虚假评论之间隐含着各种关联，可以通过挖掘二者之间的关联构建评论—评论图来研究评论之间的依赖关系。因此，可以考虑使用图上学习的神经网络模型来提取深层关系特征。

图神经网络是一种基于深度学习、图域推理的新型关系特征提取网络[37]。作为新的研究热点，GNN 在节点分类、链接预测和图分类等多个领域都取得了良好的实验效果。现在常用的基于 GNN 的图嵌入方法有图卷积网络[38]、图采样聚合网络 GraphSAGE[39]、图注意力网络(Graph Attention Networks，GAT)[40]以及基于 GNN 的变体或改进方法。Hamilton 等[39]提出的 GraphSAGE 模型由于其具备的邻居采样机制和信息聚合方式，一方面能够以相对较快的速度识别得到可信的结果，一方面扩展性能较好，因此本文选择该模型进行虚假评论预测。

将评论与评论之间的联系映射成一个共同评论图，每个评论都表示为图上的一个节点。若两个评论之间存在以下 3 种条件中的一种：①由同一评论者发布；②在同一产品评论页中，且评分相同；③在同一产品评论页中，且为同年同月发表，则用一条无向边将其连接起来，形成评论网络。将评论网络建模为图 G=(V,E)，其中 V={v_1, v_2, ..., v_n}为评论集合，E 为连接着评论的无向边集合。

由可疑值算法可知，存在 3 个影响网络中节点可疑值的因素：①初始可疑值；②与当前节点相连的其他节点的可疑值大小；③与当前节点相连的其他节点的度大小。为此，在采样时将优先采样可疑值高的节点，对于度值较小的节点，采样其二阶邻居的节点来补充邻域数量、扩大感受野，增强模型学习结构信息的能力。

使用评论网络节点的平均度值作为衡量度值大小的划分标准，计算公式如下：

$$\bar{d}=\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}A_{i,j}}{N}\quad(12)$$

其中，N 是节点总数量，$A_{i,j}$ 是邻接矩阵第 i 行第 j 列的元素值。因为本文构建的评论网络为无权无向图，所以节点 i 与节点 j 相连时值为 1，不相连时值为 0。

在直接邻接点数量不足的情况下，对可疑值高的二阶邻接点进行采样。使用自适应生成邻域的采样方法进行采样：

(1)首先输入评论网络 $\boldsymbol{G}$、所有节点的直接邻接点 $N(V)$、节点的二阶邻接点 $SN(V)$、网络平均度值 $\bar{d}$、固定的采样数量 S、节点度值 $degree(u)$、目标节点 u 的直接邻接节点 $N(u)$、二阶邻接点 $SN(u)$。

(2)计算 $\boldsymbol{G}$ 中每一个节点的可疑值，得到一个可疑值列表 SV。

(3)若 $degree(u)<0.5\cdot\bar{d}$，采样结果集合 $N^{ss}(u)=N(u)\cup SN(u)$。

(4)若 $degree(u)<0.5\cdot\bar{d}$ & $degree(u)<\bar{d}$，最终采样结果集合 $N^{ss}(u)=N(u)\cup TOPN(SN(u))$，$TOPN(SN(u))$ 为可疑值最高的 N 个二阶邻接点。

(5)若 $degree(u) \geqslant \bar{d}$ & $degree(u) < 2\cdot\bar{d}$,最终采样结果集合 $N^{ss}(u) = N(u)$。

(6)若 $degree(u) \geqslant 2\cdot\bar{d}$,最终采样结果集合 $N^{ss}(u) = TOPN(N(u))$,$TopN(N(u))$为可疑值最高的 N 个邻接点。

(7)对节点集 V 中每一个节点 v 都执行步骤(3)—(6),得到每一个节点对应的邻域采样列表 $N^{ss} = \{N^{ss}(v_1), N^{ss}(v_2), \cdots, N^{ss}(v_{|V|})\}$。

结合可疑值改进自适应生成邻域采样方法后的 TR-GraphSAGE 算法流程如下:

算法 2:TR-GraphSAGE 算法

输入:评论网络 $\boldsymbol{G}$,迭代次数 num_epochs,领域采样列表 N^{ss}

输出:节点的嵌入向量 ***embedding***

1.FOR node IN $\boldsymbol{G}$:

2.　　embedding[node] = random initialization

3.FOR i IN range(num_epochs):

4.　　FOR node IN $\boldsymbol{G}$:

5.　　sampled_neighbors = (N^{ss}, num samples)(Nss, num samples)

6.　　aggregator_input = concatenate (embedding[sampled_neighbors])

7. embedding[node] = aggregator(aggregator_input)

8.RETURN ***embedding***

TR-GraphSAGE 算法流程如下:首先在第一行和第二行初始化节点的嵌入向量,然后从第三行开始迭代更新节点的嵌入向量,第五行根据可疑值算法得到的邻域采样列表采样邻居节点,第七行聚合邻居的嵌入向量,最后在第八行返回节点的嵌入向量。使用可疑值算法后,节点的可疑值表示该节点被怀疑的概率,可疑值越高的节点被采样的概率越大,使得自适应生成邻域的采样方法采样的节点集合包含更为可疑的邻居节点,提高节点的嵌入表示质量。同时尽可能消除对虚假检测无贡献的噪音,从而保证邻域质量。

3　实验结果分析

3.1　评价标准

针对虚假评论检测这种二分类任务,使用机器学习度量指标中的精确率、召回率、F1 值作为度量指标。

精确率(Precision):用于衡量模型预测的准确性,其定义是正确预测为虚假评论与所有预测为虚假评论之比,如公式(13)所示。式中,TP 表示被正确预测的评论数,FP 表示被错误预测为真实评论的数量。

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

召回率(Recall):用于衡量模型检测正例的性能,其定义是正确预测为虚假评论与所有虚假评论之比,如公式(14)所示。其中,FN 表示被错误预测为虚假评论的数量。

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

F1 为基于精确率和召回率的综合评价指标,如公式(15)所示:

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{15}$$

3.2　特征重要性分数计算

3.2.1　数据归一化

使用线性归一化方法将原特征数值映射到[0,1]区间,同时使映射后的数值保持一致的分布。对于特征 $f_n = \{f_{n,1}, f_{n,2}, \ldots, f_{n,m}\}$,其中 $f_{n,m}$ 表示第 n 个特征在样本空间第 m 个样本中的值。归一化公式如下:

$$f_{n,m} = (f_{n,m} - f_{n,min})/(f_{n,max} - f_{n,min}) \tag{16}$$

其中,$f_{n,min}$、$f_{n,max}$ 分别表示第 n 个特征在所有样本中的最小值和最大值。

3.2.2　特征重要性分数

实验使用 Python 中的 scikit-learn 包,采用默认设置实现,并使用 train_test_split()函数随机划分样本数据集,其中 70% 的数据作为训练集,30% 的数据作为测试集。然后用 2.1.2 节中提取的特征作为输入,得到的特征重要性分数如表 1 所示。

Table 1　Input feature importance scores

表 1　输入特征重要性分数

排名	特征	重要性分数
1	赞同数量	25.876
2	好友总数量	12.639
3	总评论数量	8.010
4	积极评论比例	6.033
5	消极评论比例	5.011
6	评分偏差	4.212
7	文本相似度	3.308
8	最大发布评论数量	2.913
9	活跃时间窗口	2.510
10	创建天数	1.756
11	平均发布比例	1.127
12	文本长度	0.844

由表 1 的特征排名显示,赞同数量和好友数量是所使用数据集中最重要的特征,说明社交属性非常有区分度,在分类中贡献巨大。虚假账号没有经营社交网络的意愿,所以其只有少量好友或者无好友,而真实账号则有完整的社交圈。表中积极评论比例的贡献大于消极评论比例,该结果支持了虚假评论有 80% 都是积极评论的观点。特征表现中与时间有关的特征表现均排在最后几名,说明虚假账号与真实账号在活跃时间和创建时间上可能并不存在太大差异。

3.3　改进图神经网络实验结果

3.3.1　实验设计

改进后的图神经网络命名为 TR-GraphSAGE,其模型使用两层 TR-GraphSAGE 网络,第一层 TR-GraphSAGE 网

络的采样邻接点数量为25，第二层TR-GraphSAGE网络的采样邻接点数量为10，最后TR-GraphSAGE模型将评论—评论网络转换为100维的节点网络嵌入。

为了探索Tr-GRAGE对虚假评论检测的影响，验证Tr-GRAGE的有效性，实验选取LSTM、TextCNN、GCN和GraphSAGE 4种方法进行比较。其中，LSTM是基于深度学习的模型，TextCNN是常用于文本分类的模型，而GCN和GraphSAGE是基于图神经网络的网络嵌入模型。

3.3.2 实验结果与分析

实验使用精确率、召回率、F1值作为度量指标对模型性能进行评估，训练过程如图2所示。由实验结果可知，在基于Yelp餐厅评论数据集的真假评论分类任务中，TR-GraphSAGE模型有着较好的分类效果，其准确率、召回率和F1最高可达0.79、0.92与0.85。

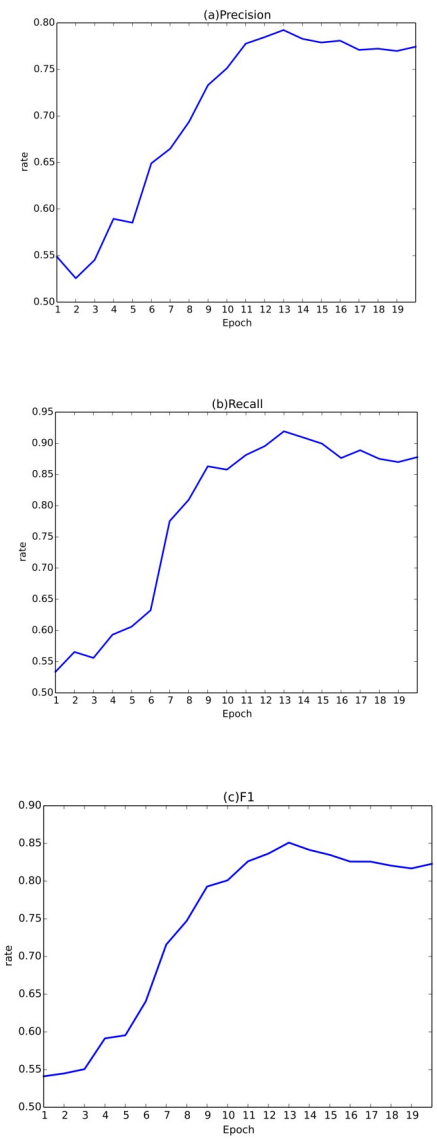

Fig. 2 Training process of TR-GraphSAGE on the dataset

图2 TR-GraphSAGE在数据集上的训练过程

对比实验结果如表2所示，其中Improving表示TR-GraphSAGE相比于其他4种模型的平均提升程度。由对比实验可知，TR-GraphSAGE模型在保证精确率不下降的同时，可实现比其他网络嵌入模型更好的查全效果。基于图神经网络的模型相比基于深度学习的模型在分类效果上表现更为突出，说明图神经网络能够从数据本身和数据之间关联两个角度进行建模，为预测提供更多特征和结构信息，具有更强的普适性。TR-GraphSAGE模型相比基于传统采样聚合式图神经网络的GraphSAGE模型在寻找潜在虚假评论方面的性能有了更大提升，召回率提高了13.59%，表明使用基于TrustRank的可疑值算法进行节点采样，可以在保持领域质量和数量的情况下优先采样有助于检测虚假评论的节点，很好地降低了噪音干扰，从而提高传统GraphSAGE模型的性能。因此，所提出的TR-GraphSAGE模型在维持精确率不发生剧烈变化的同时，能够挖掘更多未被检测出来的目标。

Table 2 Performance of each model

表2 各模型表现 **(%)**

模型	Precision	Recall	F1
LSTM	62.66	73.93	67.83
TextCNN	71.29	70.14	71.71
GCN	77.22	70.21	73.53
GraphSAGE	76.45	78.33	76.72
TR-GraphSAGE	79.23	91.92	85.10
Improving	5.86	15.01	10.12

4 结语

本文基于图神经网络对虚假评论检测进行研究，采用爬取Yelp收集的真实数据集，提取虚假评论的文本特征和行为特征，计算出特征的重要性分数，以此计算评论图中节点的可疑值，优化图神经网络算法。实验结果表明，本文提出的TR-GraphSAGE与LSTM、TextCNN、GCN和GraphSAGE相比，在精确率上分别提升了16.57%、7.94%、2.01%和2.78%，在召回率上分别提升了17.99%、21.78%、21.71%和13.59%。实验证明了引入可疑值概念可以对GraphSAGE模型的噪音干扰问题进行干预。GraphSAGE本身具有良好的泛化性能，因此本文提出的方法有望灵活应用于其他技术领域。在对数据集的分析中发现，现实场景中的虚假评论数量远低于真实评论数量，极度不平衡，并且从特征的重要性分数可以看出社交属性是有显著区分度的特征，所以未来研究可从数据的类不平衡问题和考虑评论之间的社交接触两方面对检测模型进行改进。

参 考 文 献：

[1] BUDHI G S, CHIONG R, PRANATA I, et al. Predicting rating polarity through automatic classification of review texts[C]// Kuching: 2017 IEEE Conference on Big Data and Analytics, 2017.

[2] SONG W, LI W, GENG S. Effect of online product reviews on third parties' selling on retail platforms[J]. Electronic Commerce Research and

Applications, 2020, 39(1): 101-112.
[3] FELBERMAYR A, NANOPOULOS A. The role of emotions for the perceived usefulness in online customer reviews[J]. Journal of Interactive Marketing, 2016, 36(11):60-76.
[4] GOSSLING S, HALL C M, ANDERSSON A C. The manager's dilemma: a conceptualization of online review manipulation strategies[J]. Current Issues in Tourism, 2018, 21(1):484-503.
[5] BOLTON R J, HAND D J. Statistical fraud detection: a review[J]. Operations Research, 2004, 17(3): 235-255.
[6] ZHAO Y, YANG S, NARAYAN V, et al. Modeling consumer learning from online product reviews[J]. Marketing Science, 2011, 32(1): 153-169.
[7] ZHANG W, XIE R, WANG Q, et al. A novel approach for fraudulent reviewer detection based on weighted topic modelling and nearest neighbors with asymmetric Kullback - Leibler divergence[J]. Decision Support Systems, 2022, 157(6):113-128.
[8] TIAN Y, MIRZABAGHERI M, TIRANDAZI P, et al. A non-convex semi-supervised approach to opinion spam detection by ramp-one class SVM[J]. Information Processing & Management, 2020, 57(6):102381.
[9] ELMOGY A M, TARIQ U, MOHAMMED A, et al. Fake reviews detection using supervised machine learning[J]. International Journal of Advanced Computer Science and Applications, 2021,36(1):601-606.
[10] WANG Z, WANG H, HU R L, et al. Detection of fake store reviews based on supervised learning[J]. Software Guide,2020,19(4):71-74.
王琢,汪浩,胡润龙,等. 基于有监督学习的店铺类虚假评论检测[J]. 软件导刊,2020,19(4):71-74.
[11] KONTSEWAYA Y, ANTONOV E, ARTAMONOV A A. Evaluating the effectiveness of machine learning methods for spam detection[J]. Procedia Computer Science, 2021, 190(2):479-486.
[12] ZHANG D, LI W W, NIU B Z, et al. A deep learning approach for detecting fake reviewers: exploiting reviewing behavior and textual information[J]. Decision Support Systems, 2023, 166(1):113911.
[13] SUNYOUNG H, HYUNAE L, CHULMO K, et al. Fake reviews or not: exploring the relationship between time trend and online restaurant reviews[J]. Telematics and Informatics, 2021, 59(2):101560.
[14] SASTRAWAN I K, BAYUPATI I, ARSA D. Detection of fake news using deep learning CNN-RNN based methods[J]. ICT Express, 2022, 8(3):396-408.
[15] GOLDANI M H, SAFABAKHSH R, MOMTAZI S. Convolutional neural network with margin loss for fake news detection[J]. Information Processing & Management, 2021, 58(1):102418.
[16] RUAN N, DENG R, SU C. GADM: manual fake review detection for O2O commercial platforms[J]. Computers & Security, 2020, 88(1): 101657.
[17] WANDA P, HUANG J J. DeepProfile: finding fake profile in online social network using dynamic CNN[J]. Journal of Information Security and Applications, 2020, 52(6):102465.
[18] BATHLA G, SINGH P, SINGH R K, et al. Intelligent fake reviews detection based on aspect extraction and analysis using deep learning[J]. Neural Computing and Applications, 2022, 34(22):20213-20229.
[19] CHEN Y F. Fake review detection using CNN-LSTM and transfer learning[J]. Software Guide, 2012,21(2):63-67.
陈宇峰. 采用 CNN-LSTM 与迁移学习的虚假评论检测[J]. 软件导刊,2022,21(2):63-67.
[20] QIU J T, WANG S Y. A deep matching model for detecting reviews mismatched with products in e-commerce[J]. Applied Soft Computing, 2022, 129(11):109624.
[21] SIMRAN B, NIHARIKA G, SANDEEP K S. A novel user-based spam review detection[J]. Procedia Computer Science, 2017, 122: 1009-1015.
[22] RASTOGI A, MEHROTRA M, ALI S S. Effective opinion spam detection: a study on review metadata versus content[J]. Data Inform, 2020, 5(2):76-110.
[23] MARTENS D, MAALEJ W. Towards understanding and detecting fake reviews in app stores[J]. Empirical Software Engineering, 2019, 24(6): 3316-3355.
[24] AKRAM A U, KHAN H U, IQBAL S, et al. Finding rotten eggs: a review spam detection model using diverse feature sets[J]. KSII Transactions on Internet and Information Systems, 2018, 12(10):5120-5142.
[25] LI N, DU S, ZHENG H, et al. Fake reviews tell no tales? dissecting click farming in content-generated social networks[J]. China Communications, 2018, 15(4): 98-109.
[26] SHEHNEPOOR S, SALEHI M, FARAHBAKHSH R, et al. NetSpam: a network-based spam detection framework for reviews in online social media[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(8): 1585-1595.
[27] SONG C, TENG Y, WU B, et al. Dynamic graph neural network for fake news detection[J]. Neurocomputing, 2022, 505(21):362-374.
[28] NEISARI A, RUEDA L, SAAD S. Spam review detection using self-organizing maps and convolutional neural networks[J]. Computers & Security, 2021, 106(7):102274.
[29] CAO D W, LI S M, CHEN H C. Fake review detection method based on GCN [J]. Computer Engineering and Application, 2012, 58(3): 181-186.
曹东伟,李邵梅,陈鸿昶. 基于GCN的虚假评论检测方法[J]. 计算机工程与应用,2022,58(3):181-186.
[30] TANG J N, WANG Y Q, CAO J, et al. Inter- and intra-graph attention aggregation learning for multi-relational GNN spam detection[J]. Procedia Computer Science, 2022, 214:1522-1530.
[31] LIU Z, CHEN C, LI L, et al. GeniePath: graph neural networks with adaptive receptive paths[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2018:4424-4431.
[32] LI A, QIN Z, LIU R, et al. Spam review detection with graph convolutional networks[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019: 2703-2711.
[33] MUKHERJEE A, VENKATARAMAN V, LIU B, et al. What yelp fake review filter might be doing?[C]// Massachusetts: Proceedings of the International AAAI Conference on Web and Social Media, 2013.
[34] ETAIWI W, NAYMAT G. The impact of applying different preprocessing steps on review spam detection[J]. Procedia Computer Science, 2017, 113:273-279.
[35] ZHANG D, ZHOU L, KEHOE J L, et al. What online reviewer behaviors really matter? effects of verbal and nonverbal behaviors on detection of fake online reviews[J]. Journal of Management Information Systems, 2016, 33(2):456-481.
[36] GYONGYI Z, GARCIA-MOLINA H, PEDERSEN J. Combating web spam with trustrank[C]//Proceedings of the 30th International Conference on Very Large Data Bases, 2004: 576-587.
[37] BRUNA J, ZAREMBA W, SZLAM A, et al. Spectral networks and locally connected networks on graphs[C]//Banff: Proceedings of the 2nd International Conference on Learning Representations,2014.
[38] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]//Toulon: International Conference on Learning Representations, 2017.
[39] HAMILTON W L, YING R, LESKOVEC J. Inductive representation learning on large graphs[C]//California: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017.
[40] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks[C]//Vancouver: 6th International Conference on Learning Representations, 2018.

(责任编辑:黄　健)