基于深度图卷积网络的社交机器人识别方法

毛文清[1,2]，徐雅斌[1,2,3*]

(1. 北京信息科技大学网络文化与数字传播北京市重点实验室 北京 朝阳区 100101；
2. 北京信息科技大学计算机学院 北京 朝阳区 100101；
3. 北京信息科技大学大数据安全技术研究所 北京 朝阳区 100101)

【摘要】提出了一种基于深度图卷积神经网络的社交机器人识别方法。首先，在元数据特征的基础上，引入 RoBERTa 模型进行博文情绪分类，进一步提取更能区分社交机器人和普通人的情绪多样性特征；同时采用 single-pass 进行博文聚类，构造博文相似图；在此基础上，提出了在 GCNII 模型上增加 Attention 机制的 A-GCNII 模型，通过捕捉用户元数据特征和社交网络中同一话题下的用户关系结构特征识别社交机器人。在真实新浪微博数据集上进行对比实验的结果表明，该方法在识别准确性和效果上均表现良好。

关 键 词 注意力机制； 深度图卷积网络； 情绪多样性特征； 社交机器人
中图分类号 TP391 **文献标志码** A doi:10.12178/1001-0548.2021280

Social Bot Identify Method Based on Deep Graph Convolutional Network

MAO Wenqing[1,2] and XU Yabin[1,2,3*]

(1. Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science & Technology University Chaoyang Beijing 100101;
2. School of Computer, Beijing Information Science & Technology University Chaoyang Beijing 100101;
3. Big Data Security Technology Research Institute, Beijing Information Science & Technology University Chaoyang Beijing 100101)

Abstract This paper presents a method of social robot recognition. This method extracts the characteristics of account sentiment diversity and uses the RoBERTa (robustly optimized BERT pretraining approach) model to classify the sentiment of blog posts. At the same time, the single-pass method is used to cluster blog posts and construct blog similarity graph. On this basis, attention-GCNII (A-GCNII) model, which adds Attention mechanism on the basis of graph convolutional network via initial residual and identity mapping (GCNII) model, is proposed to identify social robots by capturing user metadata features and user relationship structure features under the same topic in social networks. The results of comparative experiments on real Sina Weibo datasets show that our proposed method performs well in recognition accuracy and recognition effect.

Key words attention mechanism; deep graph convolutional network; sentiment diversity feature; socialbots

社交机器人是目前活跃于社交网络上的一种虚拟机器人。它实际上是一种自动化程序，能够利用社交账号，运用人工智能等相关技术模仿人类行为在社交网络中活动。据估计，2019 年 Facebook 活跃账户中机器人的平均存在率为 11%[1]。受政治或经济利益驱动，社交机器人的数量或比例还在呈现不断增加的趋势。Twitter 用户中进行美国股市趋势预测的有 71% 可能是社交机器人[2]；且在 2020 年的新冠病毒传播预测中也有同样的额情况[3]。由此看出，社交机器人正日益成为社交网络中影响社会舆论的重要工具之一。

研究人员在检测社交机器人方面做了大量的工作。现有的社交机器人检测模型可分基于特征的方法和基于图论的方法两类。

收稿日期：2021－10－08；修回日期：2022－02－24
基金项目：国家自然科学基金 (61672101)；网络文化与数字传播北京市重点实验室开放课题 (ICCD XN004)；信息网络安全公安部重点实验室开放课题 (C18601)
作者简介：毛文清 (1997－)，女，主要从事社交网络与安全方面的研究.
*通信作者：徐雅斌，E-mail：xyb@bistu.edu.cn

1) 基于特征的社交机器人检测方法是目前主流的检测方法。它通常将机器学习算法应用于待检测的账户，以确定这些账户是社交机器人还是人类。文献 [4-6] 通过提取简单的用户特征，利用贝叶斯模型、K 近邻模型与 C5 决策树检测社交机器人。除此之外，研究学者注意到，正常用户与社交机器人账号之间在推文中所暗含的情绪因素有很大的不同[7]。文献 [8] 通过情感分析和采用其他用户特征识别新浪微博上的水军。文献 [9] 指出社交机器人可以利用 Twitter 情绪来制造适得其反的效果，利用确认偏差制造虚假趋势或改变公众意见。目前已有文献都是进行粗粒度情感划分工作，如提取博文的情感极性或情感强烈程度作为情感特征，还没有研究细粒度情感划分对社交机器人检测的影响问题。

近年来，深度学习方法应用越来越广泛。文献 [10] 将长短时记忆网络 (long short-time memory, LSTM) 首次用于网络垃圾邮件检测，检测准确率达到 95.25%。文献 [11] 利用卷积神经网络 (convolutional neural networks, CNN) 对 Twitter 文本进行检测。文献 [12] 利用残差网络 (residual network, ResNet)、双向门控循环单元 (bidirectional gated recurrent unit, BiGRU) 和注意力机制构建了一种新的深度神经网络模型 (residual network-bidirectional gated recurrent unit-attention mechanism, RGA)，实现对社交机器人的检测。文献 [13] 采用深度学习生成模型 (variational autoencoder, VAE) 自动编码和解码样本特征，通过度量解码表示与原始特征的差异进行社交机器人检测。虽然这些基于特征的方法取得了很好的效果，但是由于社交机器人对人类行为的模仿程度越来越高，两者之间的行为特征差异越来越小，社交机器人能够较容易地模仿用户行为。但它在社交网络结构上的关系难以发生变化，因此使用用户特征并结合结构关系进行检测是一个好方法。

2) 基于图论的检测方法便是通过描述社交机器人和正常用户两者不同的社交关联结构模式，将社交机器人检测问题转化为图中节点分类问题，然后用图挖掘算法来区分正常账户和社交机器人账户。目前，基于图的深度学习方法已被用于社交机器人检测，并获得较好的检测性能。文献 [14] 考虑节点特征和用户关注关系，首次采用图卷积神经网络技术检测垃圾邮件机器人。文献 [15] 提出了一种基于图注意力网络的半监督图嵌入模型，该方法通过捕捉用户特征和社交网络中用户之间的关注关系和转发关系来识别垃圾邮件机器人。文献 [16] 结合了图卷积网络 (graph convolutional network, GCN) 和循环神经网络 (recurrent neural network, RNN) 模型，对发布垃圾邮件的恶意机器人进行检测。

综上，本文针对社交机器人大多只含有单一情绪的情况，提出了情绪多样性特征。在常规特征的基础上，更加强调捕捉社交机器人与正常用户的情绪差异进行社交机器人的识别。针对 GCNII(graph convolutional network via initial residual and identity mapping) 模型使用静态的传播法则，存在无法自适应的问题，提出一种增加注意力机制的采用博文聚类方法构造初始图的 A-GCNII(attention- graph convolutional network via initial residual and identity mapping) 模型，这既可以检测出博文内容相似的来自同一批次生产的社交机器人，又可以降低数据采集工作量。通过在每个传播层加入参数化的注意力引导机制，给予与中心节点相同类别的邻居节点更强的关联强度，从而有效的提升分类结果。

1 社交机器人特征提取

1.1 元数据特征提取

元数据是描述数据本身及其环境的数据。账号的元数据可以较为全面地反映一个账号的特征，是进行社交机器人检测研究中常用的特征。典型的元数据特征如表 1 所示。

表 1 元数据特征

特征类型	特征名称	含义
账号特征	账号名称 (Name)	账号的用户名格式是否为“用户+数字”，若是则取值为1，否则为0
账号特征	账号等级 (Urank)	账号的等级数，等级越高，属于社交机器人的概率越小
账号特征	是否认证 (Verified)	账号是否被微博平台认证，若是则取值为1，否则为0
账号特征	微博数 (StatusesCount)	账号发布的微博总数
账号特征	关注数 (FollowCount)	账号关注其他社交账号的数量
账号特征	粉丝数 (FollowerCount)	账号被其他社交账号关注的数量
内容特征	发布平台 (Source)	若博文的发布平台是手机客户端，则该项赋值为0，否则为1
内容特征	地理位置 (Location)	账号是否允许使用地理位置，若是则取值为1，否则为0
内容特征	提及数均值 (MentionAvgCount)	账号发布博文中提及他人的平均数
内容特征	链接数均值 (URLAvgCount)	账号发布博文中链接的平均数
内容特征	话题标签数均值 (HashTagAvgCount)	即账号发布的博文中包含话题标签的平均数
时间特征	博文时间间隔方差 (TimeIntervalVariance)	账号的每条历史博文的发布时间间隔的方差值
传播特征	转发比 (ForwardRatio)	账号的转发博文数量占所有博文数量的比值

因此，本文在进行社交机器人检测的特征集构造时，首先分析了社交机器人账号及正常人类账号的元数据，选择表 1 中列举的特征作为社交机器人特征的一部分，这些特征可以反映出社交网络中社交机器人账号和正常人类账号行为之间的差异。表中，MentionAvgCount 为提及数均值：$\mathrm{MentionAvgCount}(u)=\sum_{i=1}^{n}\mathrm{weiboMention}(u^i)/n$；$\mathrm{weiboMention}(u^i)$表示用户 u 第 i 条微博中包含的@数；n 表示用户 u 发布的博文总数；URLAvgCount 为链接数均值：$\mathrm{URLAvgCount}(u)=\sum_{i=1}^{n}\mathrm{weiboURL}(u^i)/n$，$\mathrm{weiboURL}(u^i)$表示用户 u 第 i 条微博中包含的 URL 数；HashTagAvgCount 为话题标签数均值：$\mathrm{HashTagAvgCount}(u)=\sum_{i=1}^{n}\mathrm{weiboHashTag}(u^i)/n$，$\mathrm{weiboHashTag}(u^i)$表示用户 u 第 i 条微博中包含的话题标签数。

1.2　情绪多样性特征提取

在某一事件中，社交机器人为了实现其目的，必然要清晰表达某种观点或传播某种信息，并带有设定的情感。但跳出该话题与事件，则很少呈现其他的情感表达。即社交机器人的博文往往只含有单一类型的情感。而正常用户除关注该话题与事件外，还关注生活中方方面面的事物，其博文情感往往呈现多样性、复杂性的特点。因此，分析账号情感表达的丰富程度有助于区分正常用户和社交机器人。

为了衡量该特性，本文提出情绪多样性特征。首先对博文进行细粒度情绪分类，分为积极、愤怒、悲伤、恐惧、惊奇和无情绪 6 类，然后计算账号的情绪多样性特征。由于发布的博文大多文本较短、省略严重，采用传统的机器学习算法对博文进行情绪分类，很难准确抽取到句子中与情感表达紧密相关的特征，且以人工标注的单个词作为特征会忽略单词所处的上下文语义信息。

2018 年 Google 提出的文本预训练模型 BERT (bidirectional encoder representations from transformers) 则能够利用 transformer 模型超强的特征抽取能力来学习词语的双向编码表示，融合了上下文信息的词语编码能更好地进行情感决策。RoBERTa (a robustly optimized BERT pretraining approach) 模型作为“强力优化”版的基于 BERT 的预训练模型，通过训练时间更久、使用更大批次和使用更多数据等设计获得了更好的效果。因此，本文采用 RoBERTa 模型进行博文情绪分类任务。具体的情绪分类模型架构为取 RoBERTa 预训练模型的最后三层 embedding 向量与 cls 向量进行拼接，然后传入 linear 层得到预测结果。

情绪多样性特征提取的流程如下。

1) 用情绪分类模型训练已标注好的语料对参数进行调优，保存测试集准确率最高的模型作为最终用于预测情绪的模型；

2) 对微博文本进行预处理，包括分词、去停用词等；

3) 将预处理后的微博文本输入到情绪预测模型中，对每条博文分类得到对应的情绪；

4) 统计每个账号所有博文对应的情绪，计算该账号出现每种情绪的概率 $p1$、$p2$、$p3$、$p4$、$p5$、$p6$；

5) 根据概率值计算情绪种类数特征、辛普森多样性指数特征 (Simpson's diversity index)、香农−维纳指数特征 (Shannon Wiener index)。辛普森多样性指数和香农−维纳指数都是量化多样性的指标，可以反映数据集中有多少种不同类型，并且可以同时考虑到这些种类的个体分布之间的系统性关系，例如丰富性，差异性或均匀性。

①情绪种类数特征 (sentimentclassnumcount)：账号发布的所有博文涉及的情绪类别数量，即统计 $p1$、$p2$、$p3$、$p4$、$p5$、$p6$ 中不为 0 的数量。

②辛普森多样性指数特征：从账号发布的博文中连续两次抽样得到的博文包含同一类情绪的概率：

$$\mathrm{SimpsonDiversityIndex}(u)=1-\sum_{i=1}^{S}P_i(u)^2$$

式中，S为情绪数目；$P_i(u)$为用户 u 包含第 i 类情绪的概率值。

③香农−维纳指数特征：描述账号情绪类别的紊乱和不确定性，不确定性越高，多样性也就越高：

$$\mathrm{ShannonWienerIndex}(u)=-\sum_{i=1}^{S}(P_i(u))(\ln P_i(u))$$

式中，S为情绪数目；$P_i(u)$表示用户 u 包含第 i 类情绪的概率值。

2　博文聚类

本文通过对大量的社交机器人账号及正常人类账号的行为分析发现，由于社交机器人账号的操纵者一般具有比较明确的目的，且完全模仿人类的语言风格仍存在困难。正常用户发布的博文大多具有个人特色，表达内容各异，发布极为相似内容的博文的可能性较低。而某一话题下来自同一批次生产

的社交机器人则使用同一语言模板，博文内容相似的可能性较高。因此，将同一话题下的相似博文聚为一类有助于发现社交机器人。鉴此，本文采用博文聚类方式进行博文相似图的构造。

由于 single-pass 聚类算法是一种增量聚类算法，每条文本只需要流过算法一次，它可以很好地应用于话题监测与追踪、在线事件监测等，特别适合如微博帖子信息的流式数据。因此，本文采用 single-pass 聚类算法来完成博文聚类的任务。

综上，首先采用 single-pass 算法进行博文聚类，然后利用博文聚类的结果构造完全图，由此得到博文相似图，整体流程如下。

1) 将待分类账号在某一话题下发布的博文保存在 txt 文件中，每行对应一条博文；

2) 将 txt 文件输入 single-pass 模型中，single-pass 算法读取 txt 文件的第一条博文，建立一个主题，并加入该主题所在的簇；

3) single-pass 算法读取下一条博文，计算该条博文与当前所有主题的余弦相似度，如果相似度值大于阈值 θ，则加入该主题所在的簇；如果相似度值小于阈值 θ，则为该条博文新建一个主题，直到遍历完 txt 文件的每一条博文，结束；

4) 所有博文聚类到不同的簇，处于同一个簇的博文互为相似博文，并规定处于同一个簇的账号之间有边相连，构造完全图，由此完成博文相似图的构造。

3 社交机器人识别

3.1 问题描述

社交网络中的用户可以分为正常用户和社交机器人。假设用户集为$V=\{v_1,v_2,\cdots,v_n\}$，类别集为$C=\{C_m,C_b\}$，C_m为正常用户集，C_b为社交机器人集。社交机器人识别是一个分类问题，具体如下：

$$F(v_i,c_j)=\begin{cases}0 & v_i\in C_m\\ 1 & v_i\in C_b\end{cases}\quad 1\leqslant i\leqslant|V|,\quad j\in\{m,b\}\tag{1}$$

式中，$F(v_i,c_j)\in\{0,1\}$为二元函数，0 表示用户v_i为正常用户，1 表示用户v_i为社交机器人。

3.2 分类模型

3.2.1 GCNII 模型

文献 [17] 提出了一种图卷积网络 semi-GCN，它是一种经典的 GCN 框架，其主要思想是使用切比雪夫一阶展开近似谱卷积，使每一个卷积层仅处理一阶邻域信息，然后通过分层传播规则叠加一个个卷积层，达到多阶邻域信息传播。

针对 GCN 模型因过度平滑而具有的浅层体系结构限制，文献 [18] 设计了 GCN 模型的扩展模型 (GCNII)。它具有初始残差和恒等映射两种简单而有效的技术，可有效地缓解过度平滑的问题。

普通 GCN 模型公式为：

$$\boldsymbol{H}^{(l+1)}=\sigma\left(\tilde{\boldsymbol{P}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}\right)\tag{2}$$

式中，$\tilde{\boldsymbol{P}}=\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}$；$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}_N$；$\boldsymbol{I}$为单位矩阵；$\tilde{\boldsymbol{D}}_{ii}=\sum_j\tilde{\boldsymbol{A}}_{ij}$，$\tilde{\boldsymbol{D}}$是$\tilde{\boldsymbol{A}}$的度矩阵；$\boldsymbol{H}^{(l)}$表示第 l 层节点的特征矩阵，对于输入层$\boldsymbol{H}^{(0)}=\boldsymbol{X}$；$\boldsymbol{W}$为权重矩阵。

GCNII 模型公式为：

$$\boldsymbol{H}^{(l+1)}=\sigma\left(\left((1-\alpha_l)\tilde{\boldsymbol{P}}\boldsymbol{H}^{(l)}+\alpha_l\boldsymbol{H}^{(0)}\right)\left((1-\beta_l)\boldsymbol{I}_n+\beta_l\boldsymbol{W}^{(l)}\right)\right)\tag{3}$$

式中，简单设置α_l= 0.1 或 0.2；$\beta=\log\left(\frac{\theta}{l}+1\right)$；$\theta$为超参数。

与普通 GCN 模型相比，GCNII 模型进行了两个修改：

1) 将平滑表示$\tilde{\boldsymbol{P}}\boldsymbol{H}^{(l)}$与到第一层$\boldsymbol{H}^{(0)}$的初始残差连接相结合；

2) 在权重矩阵$\boldsymbol{W}^{(l)}$中添加一个恒等映射$\boldsymbol{I}_n$。

关于初始残差连接，GCNII 将平滑表示$\tilde{\boldsymbol{P}}\boldsymbol{H}^{(l)}$与初始表示$\boldsymbol{H}^{(0)}$连接，使得当模型堆叠了许多层时，每个节点的最终表示也都至少保留来自输入层的部分α_l输入。

关于恒等映射，通过在权重$\boldsymbol{W}^{(l)}$中添加一个单位矩阵$\boldsymbol{I}_n$，保证了深度模型至少与浅层模型准确率相同。即假设β_l足够小，模型就会忽略权重矩阵$\boldsymbol{W}^{(l)}$。

3.2.2 改进的 A-GCNII 模型

原始的 GCNII 使用的是静态，无法自适应地传播法则，无法捕捉中心节点的哪个邻居节点对于中心节点的分类贡献更大。文本聚类构造的拓扑结构将发布相似言论内容的账号彼此连接，但这些账号的行为特征却不一定相似，应赋予具有相似行为特征的账号以更高的关联强度。

因此，本方案在 GCNII 模型的基础上，更改传播法则$\tilde{\boldsymbol{P}}$，在每个传播层加入参数化的注意力引导机制，通过计算邻居节点与中心节点的相似度，学习那些邻居与中心节点的更强关联性，以权衡他们对分类目标节点的贡献度。

整体 A-GCNII 分类模型结构如图 1 所示。A-GCNII 分类模型由一层输入层、若干隐藏层和一层输出层组成，输入层以一张图为输入，经过第一层全连接层，在正向传播之前将节点特征$\boldsymbol{X}$转为低

维初始表示$\boldsymbol{H}^{(0)}$；然后经过第二层卷积层，对图中所有节点及其邻居进行一次卷积操作，并使用卷积结果更新节点；再经过激活函数到达下一层卷积层。重复这一过程，直至到达输出层。在输出层，所有节点的特征被转化为任务相关的标签，以辅助分类。

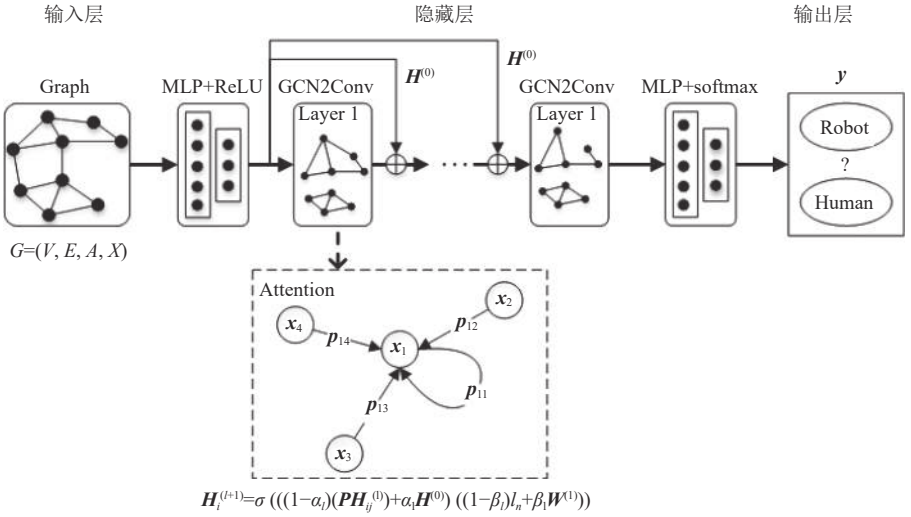


图 1　A-GCNII 结构图

图中，虚线框为第 l 层增加注意力机制后的节点聚合过程。注意力权重是通过一条边上的两个节点的特征向量的余弦相似度乘以一个自适应系数 β 后得到。每一层聚合层中共用一个 β，最后通过 softmax 使权重总和为 1。

定义传播矩阵 $\boldsymbol{P}$：若节点 i 和节点 j 之间不存在边，则元素值为 0；若节点 i 和节点 j 之间存在边，则元素值为注意力权重值，计算公式为：

$$\boldsymbol{P}_{ij}^{(l)} = \mathrm{softmax}((\beta^{(l)}\cos(\boldsymbol{H}_i^{(l)}, \boldsymbol{H}_j^{(l)}))_{j\in N(i)\cup\{i\}}) \tag{4}$$

式中，传播矩阵 $\boldsymbol{P}_{ij}^{(l)}$ 是关于第 l 层的状态和参数 $\beta^{(l)}$ 的函数。attention 中的 softmax 函数是确保传播矩阵每行的和为 1，代表邻居节点对中心节点的影响力总和为 1。从节点 j 到 i 的注意力权重计算如下：

$$\boldsymbol{P}_{ij}^{(l)} = (1/\boldsymbol{C})\mathrm{e}^{\beta^{(l)}\cos(\boldsymbol{H}_i,\boldsymbol{H}_j)} \tag{5}$$

$$\boldsymbol{C} = \sum_{j\in N(i)\cup\{i\}} \mathrm{e}^{\beta^{(l)}\cos(\boldsymbol{H}_i,\boldsymbol{H}_j)} \tag{6}$$

计算节点 i 和节点 j 在第 l 层隐含状态的余弦距离，是因为它捕捉了节点 j 到节点 i 的关联程度。注意力机制更倾向于选择那些与中心节点具有相同类别的邻居节点，并给予更强的关联强度。

由此，得到 A-GCNII 分类模型节点 i 的更新公式为：

$$\boldsymbol{H}_i^{(l+1)} = \sigma\left\{((1-\alpha_l)\left(\sum_{j\in N(i)\cup\{i\}} \boldsymbol{P}_{ij}^{(l)}\boldsymbol{H}_{ij}^{(l)}\right) + \alpha_l\boldsymbol{H}^{(0)})*((1-\beta_l)\boldsymbol{I}_n + \beta_l\boldsymbol{W}^{(l)})\right\} \tag{7}$$

4　社交机器人识别实验

4.1　数据采集与预处理

数据集由两部分数据组成：1) 通过爬虫技术爬取 2021 年 3 月 17 日−2021 年 4 月 17 日时间内微博平台上“#新疆棉花#、#我支持新疆棉花#”话题下的所有账号发布的带话题博文内容，以及爬取这些账号 ID 对应的用户信息和历史博文信息，并通过人工标注方式注明是否为机器人。经过数据预处理后，得到 6 976 个有效账号数据。2) 通过社交机器人样本数据生成模型生成机器人类型的数据。同样进行数据预处理，得到 6 636个生成账号数据。因此，本文共采用 13 612 个账号数据作为数据集，正常用户账号和社交机器人账号数量比例为 1:1，并将其按 6:2:2 划分为训练集、验证集和测试集。

4.2　评价指标

为了更真实地反映整体分类效果，本文使用准确率、精确率、查全率、F1-score 和 AUC 值 5 个常用指标来衡量提出的社交机器人检测方法的性能。

4.3 参数设置

本文使用 PyTorch Geometric(PyG) 框架，PyG 是面向几何深度学习的 PyTorch 的扩展库。处理器为 Intel® Core TM i7-10875H CPU @ 2.3 GHz，内存为 16 GB，操作系统为 Windows10。

A-GCNII 模型是基于 PyG 库的 GCNII 模型的进一步改进。模型训练时，设置层数为 8，使用学习率为 0.01 的 Adam 优化器训练模型，最多 1 000 个 epoch。设置 dropout 为 0.6，隐藏单元数量为 16，超参数 α 为 0.8，β 为 0.5，其他参数与 PyG 库中 GCNII 模型的初始参数相同。

4.4 实验

为了分析该方法检测社交机器人的有效性，设置了以下 3 组实验。实验中预设的 epoch 数为 1 000。

实验 1：情绪多样性特征有效性实验

采用 RoBERTa 模型进行博文情绪分类。首先下载 SMP2020 微博情绪分类比赛数据集，将其中的 80% 作为训练集，20% 作为测试集，将爬取的博文数据作为待分类数据。选用 RoBERTa 中文预训练语言模型作为预训练模型，得到每条博文对应的情绪类别后，提取情绪多样性特征，包含情感类别、辛普森多样性指数、香农−维纳指数。

采用 A-GCNII 分类模型来测试以下 5 种增加特征后的效果，分别为：增加 3 个情绪多样性特征 (*A*)、增加情感类别数特征 (*B*)、增加辛普森多样性指数特征 (*C*)、增加香农−维纳指数特征 (*D*)、无情绪多样性特征 (*E*)，并使用 5 个指标对于分类结果进行评价，分类评价情况如图 2 所示，且 *A*、*B*、*C*、*D*、*E* 的 AUC 值分别为：0.998 38、0.996 47、0.998 32、0.997 52，0.996 85。

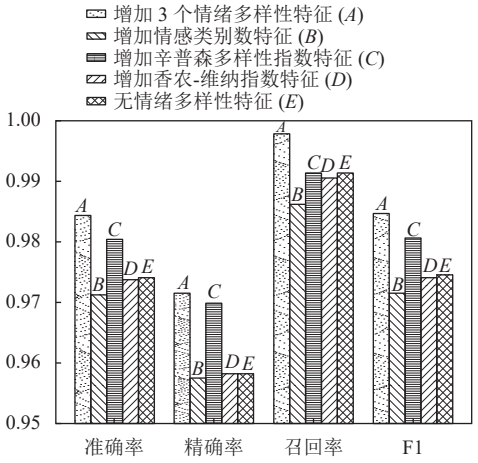


图 2 情绪多样性特征有效性验证实验结果

由图 2 可知，在 4 个指标上均呈现 *A*>*C*≥*E*≥*D*>*B* 的结果。其中，*A* 的检测效果最好，此时准确率为 98.42%，精确率为 97.13%，召回率为 99.77%，F1 值为 98.44%。由 AUC 值结果可知，虽然差异并不明显，但还是能够得出，在 ROC 曲线下面积指标上呈现 A > C ≥ D >E>B 的结果。

由此可以得出结论：增加 3 个情绪多样性特征、辛普森多样性指数特征或香农−维纳指数特征时都可以提升社交机器人的检测效果，对于预测社交网络账号是否属于社交机器人具有一定的意义。

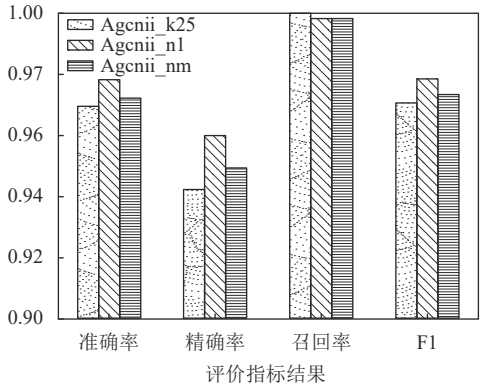
对比 3 种单一情绪多样性特征，相比于情感类别数特征，利用辛普森多样性指数特征和香农-维纳指数特征衡量情绪的多样性更有效。这是因为这两个多样性指数特征，不仅反映了数据集中包含多少种不同的类别，且考虑了这些种类的个体分布之间的系统性关系，如丰富性、差异性、均匀性。但在群落生态学分析中，辛普森多样性指数主要针对稀有种的均匀度，而香农−维纳指数针对优势种。即辛普森多样性指数更关注于社交机器人设定的针对特定话题的某一种情绪之外的其他情绪，这些情绪的数量更少，出现的可能性更小。因此利用辛普森多样性指数可以更好地凸显社交机器人账号与正常用户账号中稀有情绪的明显差异。

实验 2：博文聚类有效性实验

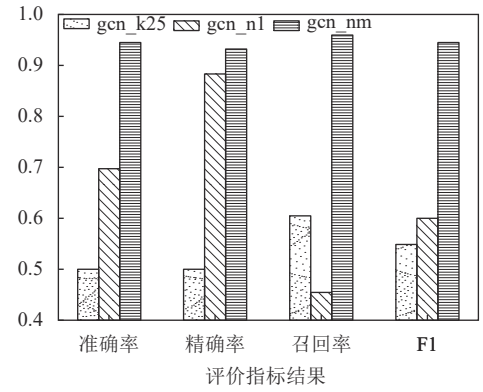
将本文提出的博文聚类构造拓扑图的方法与账号特征聚类拓扑图方法进行比较。由于本文数据集中包含部分生成数据，因此，博文聚类构造拓扑图的方法包括 nm、n1 两种具体方式。nm 表示爬取的博文经过博文聚类后聚为 *n* 类，*m* 个生成账号分为 *m* 类，总共得到 *n*+*m* 类；n1 表示爬取的博文经过博文聚类后聚为 *n* 类，*m* 条生成账号分为 1 类，总共得到 *n*+1 类。

账号特征聚类方法即对账号的特征聚类，将具有相似特征值的点聚为一类。本文采用 k-means 方法，并取 *k*=25，即表示利用 k-means 方法对爬取账号和生成账号聚为 25 类，它是利用肘方法和轮廓系数法确定的最佳聚类数。采用 GCN 和 A-GCNII 两种分类模型进行聚类检测，评价指标结果对比如图 3 所示，且 gcn_k25、gcn_n1、gcn_nm 的 AUC 值分别为：0.507 70、0.810 16、0.998 14；Agcnii_k25、Agcnii_n1、Agcnii_nm 的 AUC 值分别为：0.998 11、0.998 12、0.998 21。从图 3a 和 A-GCNII 模型的 AUC 值可看出，对于 A-GCNII 分类模型，在准确率、精确率、F1 值和 AUC4 个指标上，本文提出的博文聚类构造方式 nm 和 n1 均高

于 k-means 方法。在 recall 指标上，Recall(nm)=Recall(n1)=99.85%，Recall(k25)=1，3 种方式都呈现较高的值。其中，构造方式 n1 的检测效果最好，此时准确率为 97.83%，精确率为 95.97%，召回率为 99.85%，F1 值为 97.87%，AUC 值为 99.81%。



a. A-GCNII 分类模型的评价指标结果



b. GCN 分类模型的评价指标结果

图 3　不同拓扑结构构造方法检测结果对比图

从图 3b 和 GCN 模型的 AUC 值可看出，对于 GCN 分类模型，博文聚类构造方式 nm 在 5 个指标上的检测效果都较好，构造方式 n1 的精确率值较高，达到 88.21%，但其他 4 个指标值较低，k-means 方法在 5 个指标上的检测效果较差。综上可得，相比于 k-means 方法，两种博文聚类构造方法的检测效果更好。

比较两种博文聚类构造方法，构造方法 nm 在两个分类模型上的检测效果都很好；构造方法 n1 在 A-GCNII 分类模型上分类效果较好，但在 GCN 分类模型上分类效果较差。因此，构造方法 nm 的检测效果更稳定。进一步分析发现，由于构造方法 nm 表示 *m* 个账号发布的博文互不相同，构造方法 n1 表示 *m* 个账号发布的博文相似，显然构造方法 nm 更符合实际情况。

采用构造方法 nm 进行博文聚类，内容相似的博文聚到了第 2 107 个主题所属的类别。构造这组与博文对应账号的拓扑结构如图 4 所示，图中，三角形表示社交机器人，圆形表示正常用户，标签为对应的 ID 账号。可以看出，该组共包括 15 个账号，其中 3 个社交机器人发布了 4 条相同的博文。由此表明，采用 single-pass 聚类构造博文相似图可以挖掘出社交机器人账号间的隐秘联系，证明了构造方法 nm 的合理性与有效性。

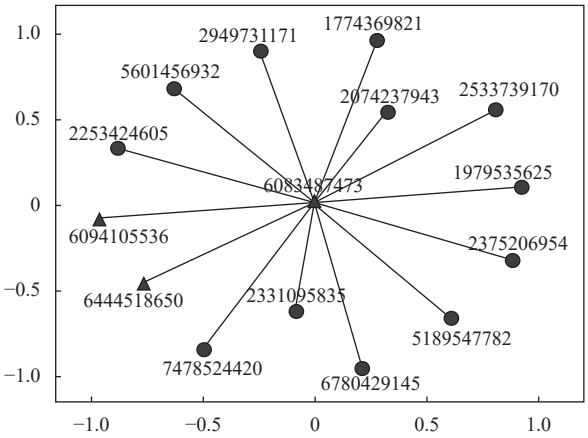


图 4　博文聚类可视化结果

实验 3：社交机器人识别对比实验

为了进一步验证 A-GCNII 模型的有效性，本文将近年来文献中直接和间接相关的模型作为基线模型，包括 3 个经典的机器学习方法、3 个深度学习方法和 3 个图神经网络方法。逻辑回归 LR[19](logistic regression)：用于新浪微博社交机器人的检测；支持向量机 SVM[8](support vector machine)：用于检测新浪微博的水军；随机森林 RF[9](random forest)：用于社交机器人检测；多层感知机 MLP[20](multilayer perceptron)：用于检测互联网水军；长短时记忆网络 LSTM[10](long short-term memory)：用于检测网络垃圾邮件；卷积神经网络 CNN[11](convolution neural network)：用于检测社交机器人；图卷积神经网络 GCN[14](graph convolutional network)：用于检测垃圾邮件机器人；图注意网络 GAT[15](graph attention network)：用于垃圾邮件机器人检测；GCNII(graph convolutional network via initial residual and identity mapping)：本文改进模型的基线方法。

选取所有 13 612 个有标签的节点，按 6:4 进行模型训练和测试，分类评价情况如表 2 所示。

表 2 分类算法实验结果对比

分类模型	准确率	精确率	召回率	F1	AUC
LR[19]	0.8912	0.8258	0.9915	0.9011	0.8912
SVM[8]	0.9498	0.9118	0.9959	0.9520	0.9498
RF[9]	0.9338	0.8832	1.0000	0.9379	0.9338
MLP[20]	0.9097	0.8476	0.9996	0.9172	0.9097
LSTM[10]	0.8971	0.8309	0.9973	0.9065	0.9908
CNN[11]	0.8352	0.8285	0.8477	0.8358	0.9211
GCN[14]	0.9364	0.9232	0.9523	0.9375	0.9831
GAT[15]	0.9303	0.9048	0.9618	0.9325	0.9818
GCNII[18]	0.9763	0.9587	0.9957	0.9768	0.9978
A-GCNII	0.9784	0.9629	0.9953	0.9788	0.9979

由表 2 可知，本文提出模型的检测效果在准确率、精确率、F1 值和 AUC 指标上均优于其他方法，在召回率指标上也接近最高值。与其他方法相比，A-GCNII 图神经网络模型的各项指标均有明显提高。对比 GCNII 模型，A-GCNII 模型的效果略有提高，这是由于引入了注意力机制，使得中心节点能够更有针对性地学习具有相似行为特征的节点特征，由此证明了 AGCNII 分类模型的有效性。

5 结 束 语

本文设计了一种结合账号情绪多样性特征的深度图卷积网络，并从账号表达情感、言论内容以及行为特征三方面对新浪微博社交机器人进行检测。通过捕捉社交机器人与正常用户在稀有情绪上的差异来更好地检测社交机器人。通过采用 single-pass 聚类构造博文相似图的方法获得图结构信息，为同一话题下的账号提供拓扑结构，降低数据采集工作量，有效地检测了来自同一批次生产的发布相似博文内容的社交机器人；最后通过在 GCNII 模型的基础上增加注意力机制，给予与中心节点相同类别的邻居节点更强的关联强度，由此提升了社交机器人的检测结果。本文在新浪微博数据集上进行实验，分析了不同特征、构图方式和分类算法对检测效果的影响。实验结果表明，本文提出的基于改进的深度图卷积网络识别模型在各个指标上均表现良好，推动了基于图的社交机器人识别的进一步发展。

参 考 文 献

[1] ZAGO M, NESPOLI P, PAPAMARTZIVANOS D, et al. Screening out social bots interference: Are there any silver bullets?[J]. IEEE Communications Magazine, 2019, 57(8): 98-104.

[2] CRESCI S, LILLO F, REGOLI D, et al. Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on twitter[J]. ACM Transactions on the Web, 2019, 13(2): 11.1-11.27.

[3] GALLOTTI R, VALLE F, CASTALDO N, et al. Assessing the risks of 'infodemics' in response to COVID-19 epidemics[J]. Nature Human Behaviour, 2020, 4(12): 1285-1293.

[4] 张艳梅, 黄莹莹, 甘世杰, 等. 基于贝叶斯模型的微博网络水军识别算法研究[J]. 通信学报, 2017, 38(1): 44-53.
ZHANG Y M, HUANG Y Y, GAN S J, et al. Weibo spammers' identification algorithm based on Bayesian model[J]. Journal on Communications, 2017, 38(1): 44-53.

[5] 谈磊, 连一峰, 陈恺. 基于复合分类模型的社交网络恶意用户识别方法[J]. 计算机应用与软件, 2012, 29(12): 1-5.
TAN L, LIAN Y F, CHEN K. Malicious users identification in social network based on composite classification model[J]. Computer Applications and Software, 2012, 29(12): 1-5.

[6] 陈侃, 陈亮, 朱培栋, 等. 基于交互行为的在线社会网络水军检测方法[J]. 通信学报, 2015, 36(7): 120-128.
CHEN K, CHEN L, ZHU P D, et al. Interaction based on method for spam detection in online social networks[J]. Journal on Communications, 2015, 36(7): 120-128.

[7] 李阳阳, 曹银浩, 杨英光, 等. 社交网络机器账号检测综述[J]. 中国电子科学研究院学报, 2021, 16(3): 209-219.
LI Y Y, CAO Y H, YANG Y G, et al. A survey of social bot detection[J]. Journal of China Academy of Electronics and Information Technology, 2021, 16(3): 209-219.

[8] JIANG Z, TROIA F D, STAMP M. Sentiment analysis for troll detection on Weibo[M]//Malware Analysis Using Artificial Intelligence and Deep Learning. Cham, Switzerland: Springer, 2021: 555-579.

[9] HEIDARI M, JONES JH, UZUNER O. An empirical study of machine learning algorithms for social media bot detection[C]//2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS). Toronto: IEEE, 2021: 656-660.

[10] MAKKAR A, KUMAR N. An efficient deep learning-based scheme for web spam detection in IoT environment[J]. Future Generation Computer Systems, 2020, 108: 467-487.

[11] ALOM Z, CARMINATI B, FERRARI E. A deep learning model for Twitter spam detection[J]. Online Social Networks and Media, 2020, 18(8): 1-12.

[12] WU Y, FANG Y, SHANG S, et al. A novel framework for detecting social bots with deep neural networks and active learning[J]. Knowledge-Based Systems, 2021, 211: 1-16.

[13] WANG X, ZHENG Q, ZHENG K, et al. Detecting social media bots with variational autoencoder and k-nearest neighbor[J]. Applied Sciences, 2021, 11(12): 1-15.

[14] ALHOSSEINI S A, Bin T R, NAJAFI P, et al. Detect me if you can: Spam bot detection using inductive representation learning[C]//Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion). San Francisco: [s.n.], 2019: 148-153.

[15] ZHAO C S, XIN Y, LI X F, et al. An attention-based graph neural network for spam bot detection in social networks[J]. Applied Sciences, 2020, 10(22): 1-15.

(下转第 629 页)