计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(7):269−275 [doi: 10.15888/j.cnki.csa.009144]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于异质图卷积注意网络的社交媒体账号分类①

陈周国[1,2], 丁建伟[1], 明　杨[3], 费高雷[3]

[1](东南大学 计算机科学与工程学院, 南京 211189)
[2](中国电子科技集团公司第三十研究所, 成都 610041)
[3](电子科技大学 信息域通信工程学院, 成都 611731)
通信作者: 陈周国, E-mail: czgexcel@163.com

摘　要: 由于社交媒体网络的复杂性, 单一性质的同质信息网络对社交媒体账号分类会造成信息丢失, 对分类结果产生不利影响. 针对这种问题, 本文提出基于异质图卷积注意网络的社交媒体账号分类方法 (HGCANA). 首先构建社交媒体的异质信息网络, 然后提取异质信息网络的社交媒体特征, 引入注意力机制, 对社交媒体账号进行分类识别. 通过实验比较 HGCANA 方法与现有方法, 证明了本文提出的 HGCANA 方法能够更好地对社交网络媒体账号进行有效分类.

关键词: 社交媒体网络; 账号分类; 异质图卷积注意网络

引用格式: 陈周国,丁建伟,明杨,费高雷.基于异质图卷积注意网络的社交媒体账号分类.计算机系统应用,2023,32(7):269–275. http://www.c-s-a.org.cn/1003-3254/9144.html

Social Media Account Classification Based on Heterogeneous Graph Convolutional Attention Network

CHEN Zhou-Guo[1,2], DING Jian-Wei[1], MING Yang[3], FEI Gao-Lei[3]

[1](School of Computer Science and Engineering, Southeast University, Nanjing 211189, China)
[2](The 30th Research Institute of China Electronics Technology Group Corporation, Chengdu 610041, China)
[3](School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract: Due to the complexity of social media networks, the classification of social media accounts by mono-nature homogeneous information networks causes information loss and has a negative impact on the classification results. To solve this problem, this study proposes a social media account classification method based on heterogeneous graph convolutional attention networks (HGCANA). Specifically, a heterogeneous information network of social media is constructed, and the social media features of the network are extracted. After that, the attention mechanism is introduced to classify and identify social media accounts. The HGCANA method is compared with the existing methods through experiments, and it is proved that the HGCANA method registers better performance in the effective classification of social media accounts.

Key words: social media networks; account classification; heterogeneous graph convolutional attention network (HGCANA)

近年来, Facebook、Twitter、YouTube、微博等社交媒体飞速发展, 将视频、音频、文字等各种信息通过社交媒体账号这个载体传播到世界各个角落, 满足人们的信息交流需求. 然而越来越多的社交媒体账号

① 收稿时间: 2022-12-13; 修改时间: 2023-01-06; 采用时间: 2023-02-03; csa 在线出版时间: 2023-04-23
CNKI 网络首发时间: 2023-04-24

利用社交媒体的便利性等, 每天都在产生海量的、重复的, 甚至是虚假的信息, 导致社交媒体平台难以有效管理, 人们也难以从“信息海洋”中直接获取有价值的信息. 为了更好地管理社交媒体用户账号, 并据此从社交媒体用户账号发表的内容中提取有价值的信息, 需要对社交媒体用户账号进行有针对性的识别分类. 社交媒体用户账号识别分类除了能够有效提升社交媒体管理系统的效率, 还能识别出相关主题的账号集合应用于问答系统[1]、推荐系统[2]、广告投放[3] 等.

本文提出的异质图卷积注意网络分类能更加全面地挖掘复杂社交媒体用户账号之间的信息, 达到了更好的账号分类识别效果. 本文剩下部分的结构如下: 在第 1 节中, 回顾社交媒体用户账号分类的相关工作. 在第 2 节中, 从异质信息网络构建、基于异质信息网络的特征提取和基于异质图卷积注意网络的账号分类 3 个方面详细的介绍本文提出的 HGCANA 方法. 在第 3 节中, 通过对比实验分析论证我们方法的性能. 最后是我们所做工作总结.

1 相关工作

信息网络被定义为一个有向网络图 $G=(V, E)$, 其中, V 是所有实体结点的集合, E 是所有关系边的集合. 并且存在着一个结点类型的映射函数 $\phi: V\rightarrow A$ 和一个边类型的映射函数 $\Psi: E\rightarrow R$, 对于每个对象 $v\in V$ 属于一种特殊的对象类型 $\phi(v)\in A$, 每个链接 $e\in E$ 属于一种特殊的关系类型 $\Psi(e)\in R$, 那么这种网络类型就是信息网络. 当对象类型的种类$|A|>1$ 或者关系类型的种类$|R|>1$ 时, 这种信息网络是异质信息网络, 否则, 它是一种同质信息网络[4]. 异质信息网络是一种信息网络, 包含了节点和边, 且节点和边具有一种或多种类型, 异质信息网络能够包含了更丰富的语义信息.

社交媒体账号分类是从海量账号及其信息中识别出具有某种特征的账号, 其分类的关键在于账号特征表示, 以及基于特征表示的分类识别算法.

1.1 账号特征表示

账号特征表示是从原始的特征中筛选出具有代表性、分类特性明显的特征信息, 主要有以下 3 种方式: 基于账号信息的表示和基于社交关系的表示. 基于账号信息的表示主要有两类, 一类是指从账号的性别、年龄、地理位置等自身属性中提取信息特征, 例如 Krishnamurth 等[5] 提取了账号的关注与被关注的比例, 而 Wang 等[6] 则是使用了好友数, 在文献 [6] 的基础上, McCord 等[7] 又增加了文本长度、关键词特征, 以及用户活跃时间分布比例等特征信息. 另一类则是指从账号发布的文本信息中提取文本长度和关键词等特征信息用于账号识别, 例如 Rao 等[8] 提取习惯用词、标点符号、表情符号等特征信息, 而 Vicente 等[9] 利用了 Twitter 账号的昵称特征. 最后一种是基于社交网络关系的特征表示, 该方法是把账号当作节点, 把关注/被关注关系、转发关系、提及关系等交互当作边, 将社交媒体数据抽象成一个社交网络图的表示方式. Pennacchiotti 等[10] 提取账号的属性、行为、文本以及社交网络特征, Campbell 等[11] 则是构建了节点为账号昵称、边为账号社交关系的带权混合图.

1.2 账号分类算法

账号分类算法是在提取到账号特征后, 根据特征信息对账号进行分类识别. 现有的账号分类方法主要是基于机器学习的分类方法, 能够自动地从输入信息数据中获取和学习特征的深度学习成为账号分类的有力工具, 但是传统机器学习方法效果非常容易受提取的特征信息质量影响[12]. Liu 等[13] 搭建了一个自动编码器的多层降噪网络对账号地理位置进行分类. Kipf 等[14] 提出图卷积神经网络 (graph convolutional network, GCN), 该网络是基于部分图形结构和节点特征进行运算, 但拥有相同邻居节点的两个不同节点往往也会有相似的特征表示. Zhang 等[15] 从账号的文本中提取出原创、转发、评论 3 种文本特征, 并使用集成的长短期记忆网络来对这 3 种不同的文本特征进行融合. Rahimi 等[16] 在 GCN 基础上, 通过分析社交媒体用户账号的文本和网络节点关系, 提出一种半监督的地理定位分类方法.

社交媒体账号分类是在数据预处理的基础上进行特征表示, 从原始的特征中筛选出具有代表性、分类特性明显的特征, 最后通过不同的算法对模型进行训练, 让模型可以对输入的数据信息进行挖掘、分析和学习, 进而对账号进行分类. 在解决社交媒体用户账号分类的问题时, 现有的方法大多是通过转发关系、好友关系、评论关系等来构建账号关系网络, 进而表征账号信息. 其方法主要是基于同质网络 (节点类型和边类型都只有一种的信息网络) 的, 但是由于社交媒体数据具有的复杂性, 使用单一性质的同质信息网络有时会直接造成重要信息的丢失, 进而对分类结果产生不

好的影响. 因此, 我们结合账号信息、交互关系信息以及账号发布的文本信息等来构建异质信息网络, 丰富网络节点类型和边类型, 以提高账号的分类准确性.

为了克服同质信息网络的不足, 本文提出了基于异质图卷积注意网络的社交媒体账号分类方法 (heterogeneous graph convolution attention network for account classification, HGCANA). 首先, 将社交媒体数据的各种信息、关系抽象成一个巨大的异质信息网络, 以反映出社交媒体网络中不同类型对象和不同类型关系之间的差异性, 并且能够描绘完整的交互信息等. 然后, 针对异质信息网络中每种节点信息的不同特点来提取不同的特征. 最后, 在卷积神经网络的基础上, 将异质信息网络嵌入其中, 并添加注意力机制为不同类型的节点分配不同的权重, 完成社交媒体账号分类. 该方法对于计算的信息类型、数量没有限制, 能够进一步挖掘社交媒体账号信息之间的隐含联系, 提高账号分类的效果.

2 基于异质图卷积注意网络账号分类

如前所述, 社交媒体用户账号分类问题的关键主要在于特征信息表示与分类算法. 现有的账号分类方法主要采用了同质网络, 在特征表示分析方面存在着特征选择单一且特征表征能力弱的缺点; 在账号信息的利用上, 由于社交媒体的复杂性和稀疏性 (即描述账号的信息很多, 但每种信息都不完整), 缺乏一种可靠的手段来对各种信息进行组织和描述. 本文采用异质信息网络融合各种类型的数据信息, 相比同质信息网络, 使用异质信息网络对账号分类带来了两方面的好处: 第一, 在信息的利用上更为全面, 异质信息网络不仅可以自然融合不同类型的对象, 还可以融入不同对象之间的交互关系, 以进一步挖掘账号信息之间的联系, 提高账号分类的准确性. 第二, 异质信息网络包含了丰富的结构特征和语义信息. 通过异质信息网络可以更加直接地发现不同类型信息之间的关系. 例如当两个账号的文本中都频繁提起某一关键词时, 即使这两个账号之间没有任何交互关系, 他们也可能是属于同一类别的账号.

虽然异质信息网络能够包含更多的信息, 但对于异质信息网络的研究还处于探索阶段中[17−19], 现有的基于异质信息网络的账号分类方法也还存在一些问题, 例如在文献 [20] 中, Dos Santos 等通过构建异质信息网络来对账号节点进行分类, 但是该方法对所有类型的节点都同等看待导致其计算复杂度极高.

本文基于异质图卷积注意网络的账号分类方法是通过对账号的社交关系、文本信息进行组织, 构建异质信息网络, 并在异质网络的基础上对账号进行分类. 其主要工作分为 3 步: 第 1 步对待分类识别数据集账号发布的文本信息进行文本预处理, 提取出需要的节点信息, 对信息进行建模后构建出异质信息网络; 第 2 步对构建的异质网络不同节点使用不同的方式提取特征; 最后一步将特征向量输入异质图卷积注意网络中, 通过融入注意力机制分析, 实现对社交媒体用户账号的分类.

2.1 异质信息网络构建

构建异质信息网络主要包含文本预处理和异质网络生成.

(1) 文本预处理

文本预处理是使用自然语言处理的方法对社交媒体数据中的文本数据进行分词、去噪、词性标注和命名实体识别. 虽然文本中含有实体等有用信息, 但也存在着许多表达不规范的内容, 因此需要对文本进行预处理. 文本预处理主要工作流程如下.

文本进行分词处理. 以文本中的空格、换行符和标点作为分隔符, 对文本进行分词处理, 将一条文本变成一个由单词组成的列表.

正则匹配去噪和提取信息. 由于列表中的单词存在着许多噪声信息, 比如表情符号、URL 链接以及一些出现频率很高但没有实际意义的单词 (例如“a”和“the”等). 这些噪声信息需要通过正则匹配来移除. 此外, 还需要通过正则匹配来提取需要的 Hashtag 信息便于后续特征的提取. 其中 Hashtag 信息是指分词后首部带有#符号的词, 比如#China.

词性标注和命名实体识别. 命名实体 (named entity recognition, NER) 是指具有特殊意义的单词, 比如人名、机构名和地名等. 命名实体通常是一些名词, 并且在这些名词的前后一般是动词或者介词, 因此需要在词性标注的基础上对命名实体进行识别.

经过上述的文本预处理流程后, 最后得到每条文本的 Hashtag 列表、命名实体列表, 以及账号之间的@提及关系列表和转发 RT 关系列表等.

(2) 异质信息网络生成

由于社交媒体信息具有复杂性和多样性的特点,

而传统的账号分类方法只考虑了信息本身对账号类别的影响, 没有考虑多种类型信息之间的隐含关系. 因此我们构建了如图 1 所示异质信息网络来将不同信息联合起来, 主要包含 3 种类型的节点.

1) 账号节点$U=(u_1,v_2,\cdots,u_w)$, 其中 w 表示账号的总个数.

2) Hashtag 节点 G, $G=(g_1,g_2,\cdots,g_x)$, 其中 x 表示 Hashtag 的总个数. Hashtag 概括了文本内容的关键词, 反映出账号对某一类事件的偏好.

3) 命名实体 NER 节点 N, $N=\left(n_1,n_2,\cdots,n_y\right)$, 其中 y 表示 NER 的总个数. 命名实体一般包含了时间、地点、人物的事件三元素信息. 从这些词语中, 可以简单了解文本的主要内容.

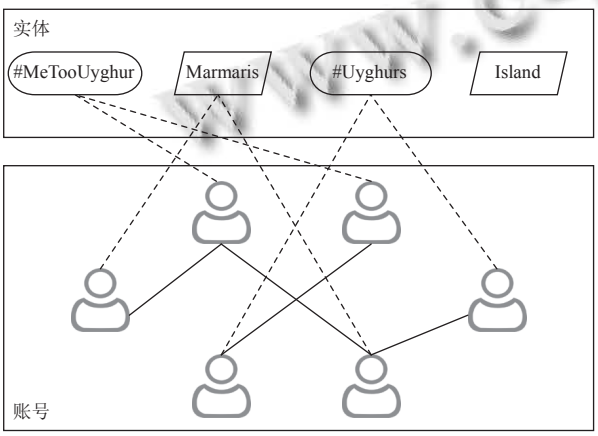

图 1 异质信息网络构建

我们并没有将账号发的文本信息直接作为节点放入异质信息网络中, 这是因为社交媒体平台对账号所发的文本有字数限定, 导致了文本的稀疏性. 但是, 账号发布的文本中又包含了许多重要的价值信息. 因此, 我们通过对文本中的内容进行关键词的提取, 将这些关键词作为我们的异质信息网络节点.

对于异质信息网络中不同节点的边连接关系, 我们将基于多种社交关系, 例如@关系列表、转发 RT 关系列表以及账号的好友关系来将账号集里的账号进行相互连接. 而对于账号和关键词 Hashtag 以及账号和命名实体 NER, 我们根据账号的文本中是否出现这些关键词, 来判断是否与账号进行连接.

2.2 基于异质信息网络的特征提取

为了将异质信息网络嵌入到后续的异质图卷积注意网络中, 我们需要针对异质信息网络中每种节点信息的不同特点来提取不同的特征.

对于账号节点, 我们采取 Node2Vec 的方法得到账号节点的特征向量. Node2Vec 综合考虑了广度优先搜索算法和深度优先搜索算法, 通过随机游走采样, 得到节点的序列组合, 既考虑同质性又考虑了同构性. 通过 Node2Vec 可以找到账号节点之间潜在的信息, 最终得到账号的特征向量$U=(k_1,k_2,\cdots,k_m)$, 其中 m 表示向量的维度.

对于 Hashtag 和 NER 节点, 由于文本的稀疏性, 会得到大量不同的 Hashtag 和 NER. 因此, 本文考虑了不同词之间的顺序与语义联系, 采用将文本中的词语进行向量化的工具 Word2Vec[21] 来提取 Hashtag 和 NER 节点的特征向量. Word2Vec 通过联系文本的上下文, 将单词投射到维度空间中成为一个向量点. 语义相似的词语出现在文本中的位置也基本相似, 因此它们投射在维度空间中的位置也会比较相近. 本文使用 Python 自带的 gensim 模块中的 Word2Vec 工具来对 Hashtag 和 NER 节点进行词向量的表征. 每一个 Hashtag 词向量表示为$g=(i_1,i_2,\cdots,i_m)$, NER 词向量表示为$n=(j_1,j_2,\cdots,j_m)$.

2.3 基于异质图卷积注意网络的账号分类

构建的异质信息网络实际上是一个不规则的图结构, 即每个节点的边连接关系不同, 连接的数量也不相同. 对这种不规则图进行处理, 只能采用 GCN 而不是传统的机器学习方法或者普通的卷积方法. 由于 GCN 只适用于同质网络, 因此我们在 GCN 的基础上提出了基于异质图卷积注意网络的账号分类方法 HGCANA, 将异质信息网络嵌入其中, 使其能适用于异质网络, 实现账号的分类.

GCN 是一个根据节点的邻居节点特征来迭代更新该节点的特征向量的多层神经网络, 如图 2 所示. 图中, 输入通道 C 中的$X\in R^{|k|A}$代表输入节点的初始特征向量矩阵, $|V|$代表节点的总个数, d表示每个节点的特征向量维度. 特征集 F 中的 Z 为输出的特征向量矩阵, Y 为标签. GCN 中不同层之间的更新为:

$$H^{(l+1)}=\sigma\left(\hat{A}H^{(l)}W^{(l)}\right) \tag{1}$$

其中, $H^{(l)}$表示第 l 层节点的特征向量, 默认第 1 层节点的特征向量是输入的特征矩阵, 即$H^{(0)}=X$. $W^{(l)}$是通过训练得到的权重变化矩阵. σ代表激活函数, $\hat{A}$为:

$$\hat{A}=\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2} \tag{2}$$

$$\tilde{A} = A + I_{|V|} \tag{3}$$

其中, $\tilde{D}$是度矩阵, A是邻接矩阵, 而$\hat{A}$是归一化后的包含了自连接的邻接矩阵, 保证每个元素的取值都在 (0, 1) 范围之间.

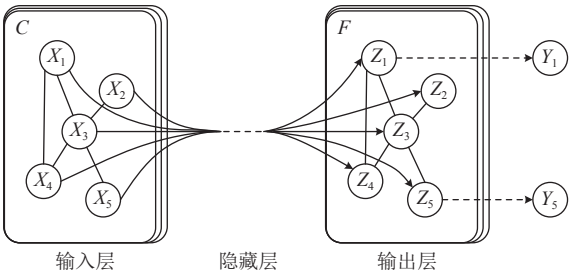

图 2 GCN 模型原理示意

HGCANA 方法在 GCN 的基础上将异质信息网络嵌入其中, 使其来对账号进行分类. 其具体步骤如下.

第 1 步, 为了更加充分地利用好账号节点、Hashtag 节点和 NER 节点信息, 对 GCN 进行改进, 使其能用于异质网络. 改进后的层之间更新公式为:

$$H^{(l+1)} = \sigma\left(ave_{i\in T}\left(\hat{A}_t H_t^{(l)} W_t^{(l)}\right)\right) \tag{4}$$

其中, T代表不同的节点类型, $A_t \in R^{|V|\times|V_t|}$是$\hat{A}$的子矩阵, 它的行代表所有的节点, 列代表该节点的邻居节点中类型为 t 的邻居节点. H_t代表不同类型节点的特征向量, W_t是不同类型节点的权重变化矩阵, 代表了输入特征与输出特征之间的关系. $ave(\cdot)$表示求平均值. 在改进后的网络中进行信息传输时, 充分利用了不同类型的节点信息, 并将不同类型节点的特征进行了融合.

第 2 步, 注意力机制的添加. 为了克服 GCN 中所有邻居节点对该节点的影响是一样大的缺点, 本文提出的 HGCANA 方法中添加了注意力机制, 从而将不同类型的邻居节点对该节点的不同影响考虑其中. 添加注意力机制的具体步骤如下.

首先, 需要区分不同类型的节点, 并给这些节点赋予不同的权重. 给定一个节点v, 用N_v^t表示节点v的 t 类型邻居节点的集合, 用e_v^t表示节点 v 的 t 类型邻居节点对节点 v 的影响分数, 具体如下:

$$e_v^t = f\left(W_t h_v, W_t h_v^t\right) \tag{5}$$

$$h_v^t = \sum\nolimits_{u\in N_v^t} \hat{A} h_u \tag{6}$$

其中, h_v表示节点v的特征向量, h_v^t是将节点v的 t 类型邻居节点的特征向量相加得到的特征向量, $f(\cdot)$是用于计算两个节点之间影响分值的函数:

$$f\left(W_t h_v, W_t h_v^t\right) = \sigma\left(\mu_t^{\mathrm{T}}\left[W_t h_v \| W_t h_v^t\right]\right) \tag{7}$$

其中, ||表示直接将两个特征向量进行横向拼接, μ_t^{T}是μ_t的转置, 是神经网络层与层之间的权重矩阵. 为了使分类效果更好, 需要对节点影响分数e_v^t进行归一化, 得到α_v^t注意力分值:

$$\alpha_v^t = \frac{\exp(e_v^t)}{\sum_{u\in N_v^t}\exp(e_u^t)} \tag{8}$$

有了不同类型节点的影响分数后, 根据式 (4) 和式 (8), 得到最终不同层之间的传递公式:

$$H^{(l+l)} = \sigma\left(agg_{t\in T}\left(a_v^t H_t^{(l)} W_t^{(l)}\right)\right) \tag{9}$$

第 3 步, 对异质图卷积注意网络进行训练. 经过 l 层的异质图卷积注意网络后, 得到节点的最终特征向量$[z_1, z_1, \cdots z_d]$, 这个特征向量的维度和账号分类的类别数相同, 里面的每个元素表示该节点属于不同分类的概率, 例如z_i表示该节点属于 i 类的概率. 最后将这个特征向量输入 Softmax 函数进行最后的分类, 具体为:

$$p_i = \frac{\exp(z_i)}{\sum_{k=1}^{C}\exp(z_k)} \tag{10}$$

其中, C 代表账号最终分类的类别数, p_i是经过 Softmax 计算后, 该节点属于 i 类别的概率. 由于使用神经网络分类是一个不断进行迭代更新, 让计算机得以学习的过程. 因此第 1 次得到的结果往往是不准确的, 因此得到预测结果后, 需要通过交叉熵损失函数与梯度下降方法, 来对权重矩阵W进行更新, 具体为:

$$loss = -\sum_{i=1}^{N}\sum_{j=1}^{c} y_{ij}\log\left(p_{ij}\right) \tag{11}$$

其中, y_{ij}是人工标记的节点 i 属于节点 j 类别的概率, p_{ij}是通过算法进行预测得到的节点 i 属于节点 j 类别的概率, N 代表输入的账号节点总个数.

3 实验与分析

3.1 数据来源

本文通过 Twitter 官方提供的 API 接口采集账号和推文等数据, 利用一些热点主题关键词从社交媒体中采集了 56 283 个账号, 随机标注 2 000 个账号用于实现特定关注账号和非关注账号的二分类任务. 为了避免过拟合现象, 按照 6:2:2 的比例将标记的 2 000 个账号划分为训练集、测试集和验证集, 当模型在训练集出现准确率上升, 而验证集的准确率却在下降的现象

时, 则停止训练.

3.2 实验结果与分析

本节将 HGCANA 方法和传统机器学习方法分类以及 GCN 方法分类作对比来证明该方法的有效性. 其中, HGCANA 和 GCN 在实验过程中有神经网络迭代的次数、dropout 的比例、隐藏层的神经元个数、神经网络的层数等参数需要调整. 选择的参数的不同, 会直接影响到最终分类效果的好坏. 这些参数都可通过实验来确定, 选取最优值. 一般来说, 神经网络的层数越多, 模型就越复杂, 学习的能力也更强. 但是这样的模型也容易出现过拟合的问题, 通常, GCN 的层数大都选择为 2 层. 实际上我们需要调整的参数主要为迭代次数和 dropout 比例.

关于迭代次数, 我们通过实验发现当迭代次数超过 400 时, 虽然训练集的准确率还在缓慢增加, 但是测试集和验证集却处于一种波动下降的状态. 这说明迭代 400 次以后, 训练的模型在训练集上可以取得更好的效果, 但是在验证集、测试集这种其他数据集上的效果反而变得不好, 说明模型的泛化能力在下降, 出现了过拟合现象.

关于 dropout 比例, 通过实验发现随着 dropout 比例的增加, 测试集的准确率处于一个先上升后下降的情况. 当 dropout 比例小于 0.5 时, 未使用的神经元比例较少, 神经元使用的个数越多, 模型越容易处于一个过拟合的状态. 因此, 当 dropout 比例增加时, 过拟合状态慢慢得到缓解, 准确率也处于一个上升的状态. 但随着 dropout 比例进一步提升, 未用的神经元个数越来越多, 导致捕捉数据特征的能力下降, 因此准确率又处于一个下降的状态.

本文使用账号分类性能评估最常用的指标准确率 (accuracy) 和 *F*1-Score 对算法的性能进行衡量. 实验结果如图 3 所示, 从图中可以发现本文提出的方法有着更高的准确率和 *F*1-Score, 可以有效提升账号分类性能. 这是因为特征提取是传统机器学习方法的瓶颈, 特征提取的好坏会直接影响到最终分类结果的好坏, 而本文基于神经网络的深度学习方法会将初始输入的特征进行组合、层级变化, 形成更加抽象复杂的数据表示, 使得计算机可以对特征进行自主学习.

最后, 本文对 GCN、HGCNA 以及 HGCANA 这 3 种基于神经网络的分类方法进行性能对比. 实验结果如图 4 所示. 从图中我们可以发现 HGCANA 的分类性能更优, 其次是 HGCNA 的分类性能, 最后是 GCN. 基于异质网络的 HGCNA 和 HGCANA 的方法要优于基于同质网络的 GCN, 主要是因为同质网络只考虑了社交媒体账号的单一信息, 单一信息的局限性使每种信息都不完整, 导致社交媒体多种信息使用不充分, 影响最终分类效果. 同时, 较之 HGCNA 的 HGCANA 拥有更优的分类性能是因为每种类型信息拥有不同的权重, 不同类型的邻居节点产生的影响大小不同, 例如同种类型的节点相比不同种类型的节点可能会带来更多的有用信息, 那相应的, 同类型节点的影响权重就应该大一些. 通过实验, 也证明了添加注意力机制的 HGCANA 在社交账号分类中更有效果.

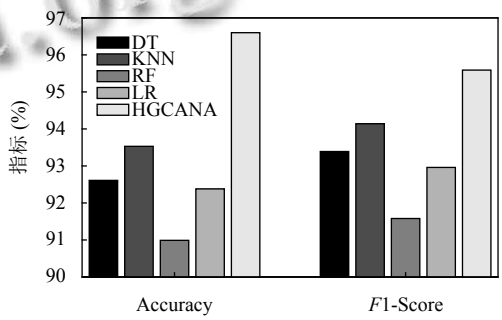

图 3 HGCANA 与传统方法的性能对比

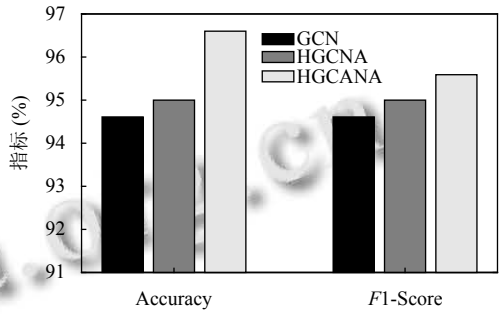

图 4 GCN、HGCNA 以及 HGCANA 的性能对比

4 总结

本文为了解决基于同质网络的社交媒体用户账号分类的局限, 提出了基于异质图卷积注意网络的账号分类方法. 首先, 我们通过账号选取、账号文本预处理和节点信息提取构建了异质信息网络. 基于构建好的异质网络, 使用不同的方法对账号节点、Hashtag 节点和 NER 节点的特征进行提取, 改进了 GCN 算法, 添加了注意力机制为不同的节点赋予不同权重以对账号进行分类. 最后, 通过实验比较 HGCANA 方法与现有社交账号分类方法, 证明了本文提出的 HGCANA 方法能够更好地对社交网络媒体账号进行有效分类.

参考文献

1 Gunawan AAS, Mulyono PR, Budiharto W. Indonesian question answering system for solving arithmetic word problems on intelligent humanoid robot. Procedia Computer Science, 2018, 135: 719–726. [doi: 10.1016/j.procs.2018.08.213]

2 Li J, Xu WT, Wan WB, *et al*. Movie recommendation based on bridging movie feature and user interest. Journal of Computational Science, 2018, 26: 128–134. [doi: 10.1016/j.jocs.2018.03.009]

3 Schreiner T, Rese A, Baier D. Multichannel personalization: Identifying consumer preferences for product recommendations in advertisements across different media channels. Journal of Retailing and Consumer Services, 2019, 48: 87–99. [doi: 10.1016/j.jretconser.2019.02.010]

4 Han J. Mining heterogeneous information networks: The next frontier. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing: ACM, 2012. 2–3.

5 Krishnamurthy B, Gill P, Arlitt M. A few chirps about Twitter. Proceedings of the 1st Workshop on Online Social Networks. Seattle: ACM, 2008. 19–24.

6 Wang AH. Don't follow me: Spam detection in Twitter. Proceedings of the 2010 International Conference on Security and Cryptography (SECRYPT). Athens: IEEE, 2010. 1–10.

7 McCord M, Chuah M. Spam detection on Twitter using traditional classifiers. Proceedings of the 8th International Conference on Autonomic and Trusted Computing. Banff: Springer, 2011. 175–186.

8 Rao D, Yarowsky D, Shreevats A, *et al*. Classifying latent user attributes in Twitter. Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents. Toronto: ACM, 2010. 37–44.

9 Vicente M, Batista F, Carvalho JP. Twitter gender classification using user unstructured information. Proceedings of the 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Istanbul: IEEE, 2015. 1–7.

10 Pennacchiotti M, Popescu AM. A machine learning approach to Twitter user classification. Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Barcelona: AAAI, 2011. 281–288.

11 Campbell W, Baseman E, Greenfield K. Content+context=classification: Examining the roles of social interactions and linguist content in Twitter user classification. Proceedings of the 2nd Workshop on Natural Language Processing for Social Media (SocialNLP). Dublin: Association for Computational Linguistics and Dublin City University, 2014. 59–65.

12 顾杰. 社交网络账号的智能分类方法 [硕士学位论文]. 成都: 电子科技大学, 2019. 3–5.

13 Liu J, Inkpen D. Estimating user location in social media with stacked denoising auto-encoders. Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. Denver: Association for Computational Linguistics, 2015. 201–210.

14 Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017. 198–212.

15 Zhang D, Li SS, Wang HL, *et al*. User classification with multiple textual perspectives. Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Osaka: The COLING 2016 Organizing Committee, 2016. 2112–2121.

16 Rahimi A, Cohn T, Baldwin T. Semi-supervised user geolocation via graph convolutional networks. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne: Association for Computational Linguistics, 2018. 2009–2019.

17 Hu LM, Yang TC, Shi C, *et al*. Heterogeneous graph attention networks for semi-supervised short text classification. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 4821–4830.

18 Ji M, Sun YZ, Danilevsky M, *et al*. Graph regularized transductive classification on heterogeneous information networks. Proceedings of the 2010 Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Barcelona: Springer, 2010. 570–586.

19 Rossi RG, de Paulo Faleiros T, de Andrade Lopes A, *et al*. Inductive model generation for text categorization using a bipartite heterogeneous network. Proceedings of the 12th IEEE International Conference on Data Mining. Brussels: IEEE, 2012. 1086–1091.

20 Dos Santos L, Piwowarski B, Gallinari P. Multilabel classification on heterogeneous graphs with Gaussian embeddings. Proceedings of the 2016 Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Riva del Garda: Springer, 2016. 606–622.

21 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. Proceedings of the 1st International Conference on Learning Representations. Scottsdale, 2013. 1–12.

(校对责编: 牛欣悦)