一种基于多阶邻居的网络环境下多标签分类算法

王 浩,张 赞,李 磊,汪 萌
(合肥工业大学计算机与信息学院,安徽合肥 230009)

摘 要: 随着标签分类应用的增长,社交网络环境下多标签分类已成为一个重要的数据挖掘研究领域. 关系分类模型基于一阶邻居做标签分类,其性能优于传统的多标签分类器. 但现有的关系分类模型也存在问题:第一,仅利用一阶邻居做分类,未能充分使用邻居信息. 第二,网络数据通常包含大量不连通的孤立部分,其标签无法利用现有的关系分类模型分类. 考虑基于共引规则为非孤立节点挖掘二阶邻居和基于节点特征向量相似度为孤立节点挖掘高阶邻居,本文提出一种新的基于多阶邻居的网络数据多标签分类算法,称为 MORN 算法. 在多个真实数据集上将 MORN 与现有的关系分类模型作对比,实验表明,MORN 算法能够学习到更多节点的标签且精度优于传统关系分类方法.

关键词: 社交网络; 关系学习; 多标签分类
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112 (2016)10-2330-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.10.007

A Multi-Order Neighbor Based Multi-Label Classification Algorithm in Network Environments

WANG Hao, ZHANG Zan, LI Lei, WANG Meng
(*School of Computer and Information, Hefei University of Technology, Hefei, Anhui* 230009, *China*)

Abstract: Multi-label classification in network environments is becoming a key area of data mining research as its applications are increasing dramatically. Relational classification models, which predict class labels of linked neighbors according to the ones of the given nodes, have been shown to outperform traditional multi-label classifiers. However, existing relational classification models neither make full use of neighbor information, nor predict the isolated nodes' labels, which are popularly existing in relational networks. In this paper, we present a multi-label relational classifier (MORN) that mines both second-order neighbors for non-isolated nodes and high-order neighbors for isolated nodes. MORN has been conducted on real datasets and it demonstrates that our proposed classifier outperforms existing relational classification models.

Key words: social networks; relational learning; multi-label classification

1 引言

近年来,随着社交网络(social networks)的发展和相关领域应用的不断深入,网络数据分类(networked data classification)已成为数据挖掘(data mining)的一个重要研究方向. 区别于传统数据,网络数据包含相互联系的实体,不遵循实例独立性假设. 实例和实例之间可以被多种多样的方式连接到一起. 本文研究基于网络数据的多标签分类问题.

传统的关系分类模型[1]基于社交网络同质性及一阶马尔科夫假设,通过群体分类(collective classification)[2],可以分类网络数据的标签. 其基本假设是:一个节点的标签依赖于网络中与它直接相连的邻居的标签[3]. 因此关系分类模型处理网络数据优于传统标签分类模型[4~6]. 但传统关系分类模型也存在两个问题:

(1)传统关系分类模型仅依赖与待分类节点直接相连的邻居节点,往往因为邻居中训练节点占比较少,影响分类精度.

(2)关系网络通常包含大量的孤立部分. 其节点的标签无法利用现有的关系分类模型分类[7].

收稿日期:2015-01-31;修回日期:2015-07-31;责任编辑:李勇锋
基金项目:国家 973 重点基础研究发展计划(No. 2013CB329604);国家自然科学基金(No. 61070131, No. 61175051);安徽省自然科学基金(No. 1408085QF130);中央高校基本科研业务费专项(No. 2015HGCH0012)

2　相关工作

2.1　传统多标签分类方法

解决传统多标签分类问题,通常基于两大类方法:第一类是将多标签分类问题拆解成多个单标签分类问题,例如 Hullermeier 等[8]提出的基于标签对比的方法.第二类是将单标签分类方法转化为多标签分类方法.Zhang 等[9]提出的 MLKNN 算法就属此类.郑等[10]将随机游走模型引入多标签学习,提出了基于随机游走模型的多标签分类方法.

传统的多标签分类算法处理网络数据存在问题:第一,数据不满足独立同分布的假设;第二,无法有效利用实例之间连接关系;第三,通常需要使用实例的特征向量作为算法输入.综上三点,传统多标签分类方法不适合网络环境多标签分类.

2.2　网络数据多标签分类方法

Tang[11]等提出了基于社会特征提取的 EdgeCluster 算法,以解决当网络数据缺少特征向量时,难以使用传统多标签分类方法的问题.该算法的整体架构属于传统多标签分类方法.

关系分类模型基于网络中的连接,充分发掘实例之间和标签之间的依赖关系,再通过群体分类,估计所有未知标签节点的标签,经过重复迭代,最终获得一个稳定状态以最小化邻居节点间标签的不一致[12].关系分类模型代表算法有 wvRN[13]和 SCRN[3].

3　MORN 算法

3.1　二阶邻居发现

共引性(citation regularity)是社会科学研究的重要现象[14,15],相似的实体倾向于连接到相同的事物.基于共引性,如果两个节点有较多的共同邻居,我们认为这两个节点存在相似性.在多标签分类计算中,我们引入了二阶邻居.

发掘二阶邻居的方法是:

(1)查找目标节点 v_i 的直接(一阶)邻居集合 N_i^1.

(2)查找 N_i^1 中包含的所有节点的邻居节点集 N_i^2,且满足 $N_i^1 \cap N_i^2 = \varnothing$.对于节点 $u_i \in N_i^2$,记 u_i 与 N_i^1 的连接数为 $Num_{u_i - v_i}$,表示 u_i 与 v_i 的共同邻居数.

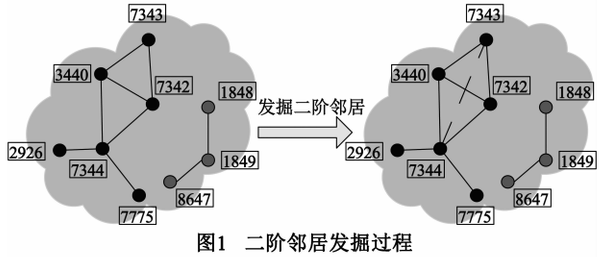

图1　二阶邻居发掘过程

(3)设定阈值 T,当 $Num_{u_i - v_i} \geqslant T$ 时,我们认为 u_i 是 v_i 的二阶邻居.

(4)我们为 u_i 和 v_i 建立新的连接,权重设为 N_i^1 所包含一阶邻居的权重最小值.

下面我们用 DBLP 数据集中一个例子来说明引入二阶邻居的必要性,该数据集的详细介绍在 4.1 节.以图 1 为例,假设 7344 号节点是待分类节点.这 5 个节点的实际标签在表 1 中.

表 1　7344 号、2926 号、3440 号、7342 和 7775 号节点的实际标签

节点	节点的标签
v_{7344}	$Y_{7344} = \{\lambda_{13}, \lambda_{14}\}$
v_{2926}	$Y_{2926} = \{\lambda_9, \lambda_{10}, \lambda_{13}\}$
v_{3440}	$Y_{3440} = \{\lambda_5, \lambda_9, \lambda_{10}, \lambda_{13}, \lambda_{14}\}$
v_{7342}	$Y_{7342} = \{\lambda_{13}, \lambda_{14}\}$
v_{7775}	$Y_{7775} = \{\lambda_{13}\}$

接下来,我们给出 wvRN 算法的模型:

wvRN 统计邻居节点具有标签 λ 的平均值 $P(Y_i = \lambda \mid v_i)$ 作为节点 v_i 具有标签 λ 的概率:

$$P(Y_i = \lambda \mid v_i) = \frac{1}{z} \sum_{v_j \in N_i^1} \omega(v_i, v_j) \times P(Y_j = \lambda \mid N_j^1) \quad (1)$$

其中 $Z = \sum_{v_j \in N_i^1} \omega(v_i, v_j)$,$\omega(v_i, v_j)$ 是 v_i 和 v_j 边的权重.

7344 号节点的一阶邻居的权重都为 1.根据式(1),7344 号节点具有各标签的概率为:

$$P(Y_{7344} \mid v_{7344}) = \begin{cases} P(Y_{7344} = \lambda_5 \mid v_{7344}) = 1/11 \\ P(Y_{7344} = \lambda_9 \mid v_{7344}) = 2/11 \\ P(Y_{7344} = \lambda_{10} \mid v_{7344}) = 2/11 \\ P(Y_{7344} = \lambda_{13} \mid v_{7344}) = 4/11 \\ P(Y_{7344} = \lambda_{14} \mid v_{7344}) = 2/11 \end{cases}$$

wvRN 默认已知 7344 号标签数目为 2.依据各标签概率(在概率相等情况下,wvRN 算法默认按标签顺序),7344 号节点的标签分类结果是:$Y_{7344} = \{\lambda_9, \lambda_{13}\}$,与实际情况不相符.

我们为 7344 号节点发掘到二阶邻居即 7343 号节点.其实际标签是 $Y_{7343} = \{\lambda_{13}, \lambda_{14}\}$.

此时为 7344 号节点和 7343 号节点建立新连接(图中虚线).使用 wvRN 算法,7344 号节点具有各标签的概率为:

$$P(Y_{7344} \mid v_{7344}) = \begin{cases} P(Y_{7344} = \lambda_5 \mid v_{7344}) = 1/13 \\ P(Y_{7344} = \lambda_9 \mid v_{7344}) = 2/13 \\ P(Y_{7344} = \lambda_{10} \mid v_{7344}) = 2/13 \\ P(Y_{7344} = \lambda_{13} \mid v_{7344}) = 5/13 \\ P(Y_{7344} = \lambda_{14} \mid v_{7344}) = 3/13 \end{cases}$$

此时,7344 号节点的标签分类结果是:$Y_{7344} =$

$\{\lambda_{13},\lambda_{14}\}$,与实际情况相符. 此例说明引入二阶邻居可以更充分地利用邻居信息以利分类.

3.2 高阶邻居发现

基于社会特征相似度,从已知标签节点集中寻找与孤立节点相似的节点集,该集合中的节点即为该孤立节点的高阶邻居节点. 相似度越大,说明孤立节点与该高阶邻居的标签集相似度越高,会被分配越高的权重.

挖掘高阶邻居的具体步骤如下:

(1)依据已知标签节点集的分布,识别出孤立节点集合 V^{isolated},具体使用的工具在 4.3 节介绍.

(2)删除孤立节点 $v_i \in V^{\text{isolated}}$ 与各一阶邻居的边,因为 v_i 的一阶邻居都是孤立节点,对标签分类不起作用.

(3)我们用 Tang 等提出的边聚类算法[11]提取所有节点的社会特征向量 SF(Social Features). 该算法在网络结构上使用聚类刻画节点的特征向量,可以在无连接的情况下使度量节点间的关系成为可能.

(4)计算孤立节点 v_i 与每一个已知标签节点 $v_j \in V^m$ 的社会特征向量的相似度 $Sim(SF_{v_i},SF_{v_j})$.

(5)我们选取与目标节点 v_i 相似度数值大于 0 的已知标签的节点作为 v_i 的高阶邻居.

(6)在网络 G 中,为每个目标节点 v_i 与其高阶邻居建立新的边连接,边的权重为归一化后的相似度值.

下面我们举例说明高阶邻居挖掘. 如图 2 所示,图中左半部分显示网络 G 由左右两个子网络 G_1 和 G_2 构成. 假设已知标签节点集 $V_m=(v_{3440},v_{7342},v_{7344})$. 我们想预测 G_2 中包含的节点的标签. 因 G_2 中不含已知标签节点,我们得出 G_2 是孤立部分,在图 2 中以灰色点表示.

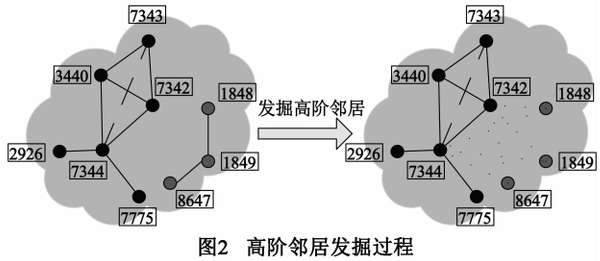

图2 高阶邻居发掘过程

在确认 v_{1848} 是孤立节点后,删除它与 v_{1849} 的连接. 计算 v_{1848} 与 V_m 中每一个节点的社会特征向量相似度. 选取相似度大于 0 的节点作为 v_{1848} 的高阶邻居. 假设发掘到的 v_{1848} 的高阶邻居是 v_{7342} 和 v_{7344}. 我们分别为节点对 (v_{1848},v_{7342}) 和 (v_{1848},v_{7344}) 建立连接边. 孤立节点 v_{1849} 和 v_{8647} 发掘高阶邻居的过程与 v_{1848} 的发掘过程一致.

综合二阶邻居发现方法和高阶邻居发现方法,重新构建网络结构和邻居关系,调用 SCRN 算法作为基础分类器给出 MORN 算法,如下所示.

算法 1 基于多阶邻居的多标签分类算法 MORN

输入:网络 G,标签集 Y,训练集 V^m,测试集 V^{test},最大迭代次数 Max_Iter

输出:V^{test} 中所有节点的标签

1. 使用文献[6]中的方法为 G 中每一个节点提取社会特征向量 $SF(v_i)$
2. 基于 V^m 的分布,计算得到孤立节点集合 V^{isolated}
3. for each $v_i \in V^{\text{test}}$ do
4. if $v_i \notin V^{\text{isolated}}$
5. 发掘 v_i 的二阶邻居 //3.1 节
6. if $v_i \in V^{\text{isolated}}$
7. 发掘 v_i 的高阶邻居 //3.2 节
8. 跟据上述 3~7 步,重新构建网络,得到 G'
9. while iterations$\leqslant Max_Iter$ and predictions have not converged to stable values
10. 基于 G',调用 SCRN 算法作为底层分类器
11. endwhile
12. 得到 V^u 中节点的标签,从中提取出 V^{test} 中节点的标签,即为测试集节点分类结果

4 实验

4.1 数据集

本文采用的 DBLP 数据集[3],提取自 DBLP 网站. IMDb 数据集[3]提取自在线视频数据库 IMDb,我们从中选取了 10000 个节点. DBLP 数据集和 IMDb 数据集统计信息如表 2 所示.

表 2 数据集统计信息

数据集名称	DBLP	IMDb
节点数目 n	8,865	10,000
边数目	12,989	295,374
标签数目 q	15	27
网络密度	3.3×10^{-4}	5.9×10^{-3}
节点最大度数	86	290
节点平均度数	3	55

4.2 度量标准

本文使用 2 个经典的标签分类度量标准 Macro-$F1$ 值和 Micro-$F1$ 值[3,8,16]度量分类精度.

本文在度量各算法分类精度的基础上,提出预测比例 PR(Prediction Ratio)指标来度量最终测试集内有多少比例的节点获得了分类标签.

$$PR=\frac{Num_{\text{predicted}}}{Num_{\text{test}}} \tag{2}$$

4.3 实验结果与分析

我们基于 Matlab 实现 MORN 算法. 采用的对比算法是 SCRN、wvRN 和 EdgeCluster.

本文使用边聚类方法提取节点的社会特征向量.

我们采用 GHI[3] 来计算特征向量相似度. 我们基于 Matlab BGL 工具包识别孤立节点. 依据文献[3]和文献[11],DBLP 数据集的社会特征向量维度设置为 1000,IMDb 数据集的社会特征向量维度设置为 10000,其它均采用默认参数.

对比实验使用网络交叉验证(Network Cross-Validation,NCV)[16] 对结果进行验证,以减小测试集的重叠.

4.3.1　二阶邻居发现中阈值 T 对实验结果的影响

在 DBLP 上做实验,比较 T 取不同值,Micro-$F1$ 的变化. 训练集规模设定为 10%,实验结果如图 3 所示. 通过实验发现,T 值取 1 时,Micro-$F1$ 值最大,为 57.02%. 当 T 值增大到 2,Micro-$F1$ 值降低到 56.59%,降幅非常明显. 随着 T 值进一步扩大,Micro-$F1$ 值变化极小. 因此,在后续实验中,我们设定 T 值为 1.

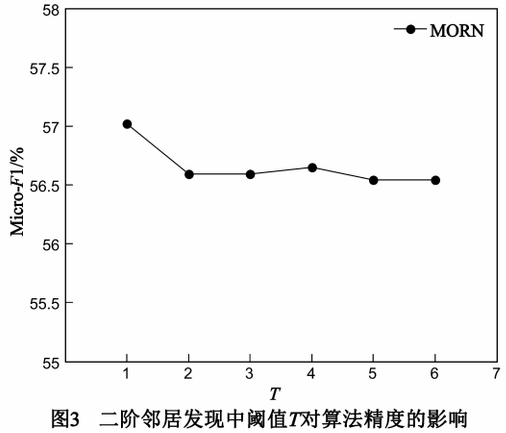

图3　二阶邻居发现中阈值T对算法精度的影响

4.3.2　MORN 算法和 SCRN 算法运算时间对比分析

我们对比两个算法在标签预测上开销的时间. MORN 算法和 SCRN 算法在标签预测阶段,迭代次数 Max_Iter 都设为 10. 由表 3 的实验结果可知,虽然 MORN 需要挖掘新的邻居,运算时间多于 SCRN,但在可接受范围之内.

表 3　MORN 算法和 SCRN 算法在 DBLP 数据集上的运算时间

训练集占数据集的比重		10%	15%	20%	25%	30%	35%
Running Time (s)	MORN	150.5	162.2	165.0	152.6	135.5	113.7
	SCRN	36.7	34.7	33.2	31.5	29.6	27.4

4.3.3　标签分类结果与分析

从表 4 和表 5 可知,MORN、SCRN 和 wvRN 的精度高于 EdgeCluster. MORN 在 DBLP 上的精度高于 SCRN 和 wvRN. 在 IMDb 上,MORN 的精度高于 wvRN,但在训练集占 1%—2% 时,MORN 的精度低于 SCRN. 原因是训练集过少,发掘到的二阶邻居和多阶邻居数量都非常少. 随着训练集增大到 3% 以上,MORN 在 IMDb 上精度好于 SCRN 和 wvRN. 对比实验说明 MORN 的精度比 SCRN 和 wvRN 更高.

表 4　各算法在 DBLP 数据集上的分类精度

训练集占数据集的比重		10%	15%	20%	25%	30%	35%
Micro-$F1$ (%)	MORN	56.65	58.52	61.35	63.36	65.47	67.84
	wvRN	53.19	56.41	58.67	60.76	62.68	64.69
	SCRN	55.78	58.10	61.14	63.12	65.28	67.25
	Edge	43.76	47.84	50.14	51.35	53.07	54.23
Macro-$F1$ (%)	MORN	50.66	52.34	54.29	57.32	59.31	61.35
	wvRN	46.47	51.26	54.32	56.57	59.55	59.83
	SCRN	48.53	52.11	54.03	57.15	59.04	60.96
	Edge	40.13	43.19	45.97	46.85	49.11	50.97

表 5　各算法在 IMDb 数据集上的分类精度

训练集占数据集的比重		1%	2%	3%	4%	5%	6%
Micro-$F1$ (%)	MORN	45.63	53.16	59.14	60.84	63.19	64.85
	wvRN	45.37	53.12	58.58	60.24	62.44	64.09
	SCRN	45.82	53.54	58.87	60.47	62.94	64.35
	Edge	39.95	46.69	51.89	54.65	57.12	59.37
Macro-$F1$ (%)	MORN	18.63	22.18	27.37	31.79	33.43	35.27
	wvRN	19.09	22.36	26.92	31.30	32.86	34.83
	SCRN	18.75	22.20	27.13	31.48	33.16	35.05
	Edge	17.52	20.67	24.39	27.18	29.81	31.34

从表 6 和表 7 可知,MORN 在 PR 值上好于 SCRN 和 wvRN,在 DBLP 上领先 13% – 15%;在 IMDb 上领先 8% – 15%. 实验结果说明,MORN 可以有效地学习到孤立节点的标签. 随着训练集规模扩大,MORN 在 PR 值上接近 100%,但没有达到 100%. 原因是少数孤立节点与所有的训练集节点的特征向量相似度都为 0,无法为其找到相似的节点.

表 6　关系分类方法在 DBLP 数据集上的预测比例

训练集占数据集的比重		10%	15%	20%	25%	30%	35%
Prediction Ratio (%)	MORN	93.58	95.17	96.24	96.63	97.43	98.47
	wvRN	78.34	79.65	83.28	83.57	84.04	85.17
	SCRN	78.34	79.65	83.28	83.57	84.04	85.17

表 7　关系分类方法在 IMDb 数据集上的预测比例

训练集占数据集的比重		1%	2%	3%	4%	5%	6%
Prediction Ratio (%)	MORN	96.23	96.45	97.34	97.63	98.27	99.22
	wvRN	81.37	83.78	85.37	86.20	87.64	90.78
	SCRN	81.37	83.78	85.37	86.20	87.64	90.78

5　结论

本文分析了现有关系分类模型在处理网络环境下多标签分类存在的问题. 针对现有关系分类模型仅利用一阶邻居做分类并且无法学习孤立节点标签的不足,提出了一种基于多阶邻居的网络数据多标签分类 MORN 算法. 理论分析和实验都证明了 MORN 算法能够有效解决网络环境下多标签分类问题.

参考文献

[1] S Macskassy, F Provost. A simple relational classifier[A]. Proceedings of the Second Workshop on Multi-Relational Data Mining at ACM SIGKDD[C]. Washington DC, USA: ACM, 2003. 64 – 76.

[2] P SEN, G Namata, M Bilgic, L Getoor, B Gallagher, T Eliassi-Rad. Collective classification in network data[J]. AI Magazine, 2008, 29(3): 93 – 106.

[3] X Wang, G Sukthankar. Multi-label relational neighbor classification using social context features[A]. Proceedings of the ACM SIGKDD[C]. Chicago, USA: ACM, 2013. 464 – 472.

[4] J Neville, D Jensen, L Friedland, M Hay. Learning relational probability trees[A]. Proceedings of the ACM SIGKDD[C]. Washington DC, USA: ACM, 2003. 625 – 630.

[5] B Taskar, P Abbeel, D Koller. Discriminative probabilistic models for relational data[A]. Proceedings of the UAI[C]. Edmonton, Canada: Morgan Kaufmann, 2002. 485 – 492.

[6] M Zhang, K Zhang. Multi-label learning by exploiting label dependency[A]. Proceedings of the ACM SIGKDD[C]. Washington DC, USA: ACM, 2010. 999 – 1007.

[7] C Aggarwal. Social Network Data Analytics[M]. Berlin: Springer, 2011. 115 – 148.

[8] E Hullermeier, J Furnkranz, W Cheng, K Brinker . Label ranking by learning pairwise preferences[J]. Artificial Intelligence, 2008, 172(16): 1897 – 1916.

[9] M Zhang, Z Zhou. A k-nearest neighbor based algorithm for multi-label classification[A]. Proceedings of the IEEE International Conference on Granular Computing[C]. Beijing, China: IEEE, 2005. 718 – 721.

[10] 郑伟,王朝坤,刘璋,王建民. 一种基于随机游走模型的多标签分类算法[J]. 计算机学报, 2010, 33(8): 1418 – 1426.
W Zheng, C Wang , Z Liu, J Wang. A multi-label classification algorithm based on random walk model[J]. Chinese Journal of Computers, 2010, 33(8): 1418 – 1426. (in Chinese)

[11] L Tang, H Liu. Scalable learning of collective behavior based on sparse social dimensions[A]. Proceedings of the ACM CIKM[C]. Hong Kong, China: ACM, 2009. 1107 – 1116.

[12] J Neville, D Jensen. Iterative classification in relational data[A]. Proceedings of the AAAI Workshop on Learning Statistical Models from Relational Data[C]. Austin, USA: AAAI, 2000. 42 – 49.

[13] S Macskassy, F Provost. Classification in networked data: a toolkit and a univariate case study[J]. Journal of Machine Learning, 2007, 8(5): 935 – 983.

[14] M Newman. Networks: An Introduction[M]. Oxford: Oxford University Press, 2010.

[15] M McPherson, L Smith-lovin, J Cook. Birds of a feather: Homophily in social networks[J]. Annual Review of Sociology, 2001, 27(1): 415 – 444.

[16] J Neville, B Gallagher, T Eliassi-Rad, T Wang, Correcting evaluation bias of relational classifiers with network cross validation[J]. Knowledge and Information Systems, 2012, 30(1): 31 – 55.

作者简介

王 浩 男,1962 年生,江苏泰州人. 合肥工业大学计算机与信息学院教授,博士生导师. 1984 年于上海交通大学获工学学士学位,1989 年和 1997 年于合肥工业大学获工学硕士和工学博士学位. 主要研究方向为智能计算理论与软件、分布式智能系统、复杂系统理论与建模等.

张 赞 男,1987 年生,安徽马鞍山人. 2009 年于东北电力大学获管理学学士学位,2013 年于合肥工业大学获工学硕士学位. 现为合肥工业大学博士研究生. 主要研究方向为社交网络行为的预测和分类研究.
E-mail: zhangzan99@163.com