应用随机游走的社交网络用户分类方法

贺超波[1,2] 杨镇雄[2] 洪少文[2] 汤 庸[2] 陈国华[2] 郑 凯[2]
(仲恺农业工程学院信息科学与技术学院 广州 510225)[1] (华南师范大学计算机学院 广州 510631)[2]

摘 要 针对现有在线社交网络用户分类方法不能有效利用用户属性和关系网络信息提高分类性能的问题,设计了一种基于随机游走模型的多标签分类方法MLCMRW。该方法的分类过程包括学习用户初始化类别标签以及通过迭代推理获得用户稳定标签分布两个阶段,并且其可以同时考虑用户属性以及关系网络特征信息进行分类。多个在线社交网络数据集上进行的实验表明,MLCMRW比其它已有的代表性方法有更好的分类性能,并且更适合对现实中的在线社交网络进行用户分类。

关键词 在线社交网络,用户分类,随机游走
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.2.042

User Classification Method in Online Social Network Using Random Walks

HE Chao-bo[1,2] YANG Zhen-xiong[2] HONG Shao-wen[2] TANG Yong[2] CHEN Guo-hua[2] ZHENG Kai[2]
(School of Information Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China)[1]
(School of Computer, South China Normal University, Guangzhou 510631, China)[2]

Abstract Aiming at the problem that the existing methods for user classification in online social network (OSN) are not enough effective to utilize both attribute and linkage information of user to improve the classification performance, we designed a new multi-label classification method using random walks (MLCMRW) to solve the problem of user classification in OSN. MLCMRW can utilize both user attribute and linkage information to improve the classification performance. In particular, MLCMRW includes two key parts: learning the initial label distribution and iterative inference for steady label distribution of every user. The experiments on the real-world OSN datasets show that MLCMRW performs quite well than other representative methods. Moreover, it is suitable to classify users in the real-world OSN.

Keywords Online social network, User classification, Random walks

1 引言

在线社交网络(Online Social Network,OSN)用户分类的本质是在确定类别范围的前提下,通过学习用户特征为用户分配所属类别标签,许多基于OSN的应用(如社会化营销、相似用户推荐以及异常用户检测等)利用用户分类的结果可以显著提高性能。目前已有不少研究人员关注OSN用户的分类问题,具有代表性的有:Z. Arkaitz等人利用用户社会标签作为特征信息源并通过训练支持向量机分类器实现用户分类[1];R. Delip等人关注于从用户Profile信息中获取用户特征信息并使用栈式支持向量机实现用户分类[2];M. Pennacchiotti等人从用户浏览历史以及交互行为记录方面分析用户特征并通过使用梯度增强决策树模型实现用户分类[3]。然而以上这些只考虑用户自身属性特征信息的方法在对现实中的OSN用户进行分类时并不能发挥很好的效果,因为它们都忽略了用户间存在的关系网络信息。OSN用户间存在清晰的联系,例如好友联系或者关注联系等。这些联系信息表明用户间是彼此依赖的,所以对于OSN用户分类问题,要执行分类的用户对象并不满足独立一致性分布的条件,因而传统的分类方法并不适合解决OSN用户分类问题。为此,Z. Wu[4],L. Tang[5]以及P. Francisco[6]等人从用户关系网络信息中学习用户特征,并分别使用SVM、集体推理(collective inference)以及PageRank等方法进行分类,在一定程度上提高了分类性能,但这些方法都只是单一使用用户的关系网络信息,事实上,用户属性信息同关系网络信息一样对用户的分类具有重要作用,在分类过程中应该同时考虑这两类信息。另外,OSN中待分类的目标用户常具有多个不一样的兴趣,例如在学者网OSN[7]上的大部分用户就具有多个研究兴趣,所以其分类结果应该有多个分类标签,但目前的用户分类方法均为单一标签分类方法。总的来说,目前OSN的用户分类仍然需要设计更好的分类方法来提高分类质量,为此本文提出一种基于随机游走模型的多标签分类方法MLCMRW用于解决

到稿日期:2014-04-14 返修日期:2014-07-09 本文受国家高技术研究发展计划(863计划)项目(2013AA01A212),国家自然科学基金(60970044,61272067,61370178),国家科技支撑计划项目(2012BAH27F05),广东省自然基金团队研究项目(S2012030006242),广东省科技计划项目(2012A080104019,2011B080100031),广东省高校优秀青年创新人才培养计划项目(2012LYM_0077)资助。

贺超波(1981-),男,博士,副教授,CCF高级会员,主要研究方向为数据挖掘与社会计算,E-mail:hechaobo@foxmail.com;**杨镇雄**(1989-),男,硕士生,主要研究方向为社交网络挖掘;**洪少文**(1989-),男,硕士生,主要研究方向为社交网络挖掘;**汤 庸**(1964-),男,博士,教授,主要研究方向为社交网络与大数据应用;**陈国华**(1984-),男,博士,讲师,主要研究方向为机器学习;**郑 凯**(1978-),男,博士,高级工程师,主要研究方向为教育信息化。

目前OSN用户分类存在的问题。所做的主要工作包括：(1)对随机游走模型尤其是图上的随机游走问题进行了研究，并给出了相关概念的形式化定义，另外对适用于图数据分类的集体分类模型进行了分析，并归纳了其中的代表性方法。(2)设计了一种基于随机游走模型的多标签分类方法MLCMRW，用于解决OSN用户分类问题，该方法的分类过程包括学习用户初始化类别标签以及迭代推理获得用户稳定标签分布两个阶段，并且可以同时考虑用户属性以及关系网络特征信息进行分类。(3)在实际的多个OSN数据集上测试多个用户分类方法的性能，实验结果表明，MLCMRW方法比其它可以应用于OSN用户分类问题的典型方法有更好的表现，并且MLCMRW方法更适合对现实中的无标度OSN进行用户分类。

本文第2节为相关工作，介绍随机游走模型、集体分类原理及代表性方法；第3节为问题定义的形式化描述；第4节介绍MLCMRW的具体实现；相关实验和分析结果在第5节介绍；最后为总结部分。

2 相关工作

2.1 随机游走模型

随机游走又称随机游动或随机漫步，在实际生活中，就存在很多与随机游走有关的现象，如醉汉的行走轨迹、股票价格的变动以及滴入水中的墨水扩散等。随机游走本质上是一种随机化描述方法，并且被认为是马尔科夫链的一种典型的表现形式，其包含的相关概念和性质如下：

随机状态变量序列 $X_0, X_1, \cdots, X_n \subseteq \Omega$ 称为具有状态空间 Ω 的马尔科夫链，如果对于所有可能的状态 $X_0, X_1, \cdots, X_n \subseteq \Omega$，有

$$\Pr[X_{t+1}=x_{t+1} \mid X_0=x_0 \wedge X_1=x_1 \wedge \cdots \wedge X_t=x_t] = \Pr[X_{t+1}=x_{t+1} \mid X_t=x_t] \quad (1)$$

那么该马尔科夫链被称为时齐的，意味着后者的概率独立于参数 t。

具有有限状态和时齐的马尔科夫链可以在称为转移概率矩阵的 P 中存储所有关于状态转移概率的信息，其中 P 为 $|\Omega| \times |\Omega|$ 矩阵，对于每一个元素 $p_{xy} \in P$，有：

$$p_{xy}=\Pr[X_1=y \mid X_0=x] \quad (2)$$

且 $p_{xy} \in [0,1]$，$\sum_y^{|\Omega|} p_{xy}=1$。随机游走过程中的每一步状态转移都可以用概率进行描述，因此非常适合于描述图节点之间的状态转移关系。不失一般性，假设存在一个无向有权图 $G=(V,E,W)$，其中 V、E 和 W 分别代表节点集合、边集合以及边权重集合，$n=|V|$ 表示节点数量，$W=[w_{ij}]n \times n$，w_{ij} 为节点 i 和 j 之间的联系边的权重，且 $w_{ij}=w_{ji}$。那么在图 G 上的一次随机游走指的是首先从某一个节点开始，然后在每一步中按照某个概率值跳转到下一个邻居节点，直至在某一个节点结束游走的过程。因为图 G 是带权图，从某个节点跳转到下一个邻居节点的概率正比于两个节点之间的边权重，其对应的概率转移矩阵为：

$$P=D^{-1}W \quad (3)$$

其中，$D=diag(d_{11}, d_{22}, \cdots, d_{nn})$，$d_{ii}=\sum_{j=1}^{n} w_{ij}$。

在图 G 上随机游走 t 步后，能获得一个概率分布矩阵 P^t，其中 P^t 中的元素代表了在随机游走第 t 步时每一个图节点被访问到的概率。一次有限步数的图上随机游走本质上是一个迭代过程，它使用上一步随机游走获得的概率分布矩阵作为下一步随机游走的初始输入概率，即：

$$P^t=P^{t-1}P \quad (4)$$

当 $t \to \infty$ 时，P^t 将可以收敛，此时可获得稳定的概率分布。文献[8]指出 P^t 稳定的概率可以用来表示图节点之间在类标签上的相似性。

2.2 集体分类

由于OSN用户之间存在关系网络，该网络适合采用图的形式进行表达，因此OSN用户分类本质上属于图数据分类问题[9]，对于该问题，集体分类方法是一种流行的解决方案。目前已有不少研究人员对集体分类进行了研究，它的主要思想是结合来自传统特征向量的监督知识和来自图结构的特征信息同时对一组关联实例进行分类，而不是独立地对每个实例进行分类。集体分类已经被广泛用于各种类型的图数据分类中，例如文献[10]对在线社交网络的用户评论进行情感分类、文献[11]对邮件通信网络进行垃圾邮件过滤以及文献[12]对蛋白质交互网络中的蛋白质功能进行类别预测等都是集体分类方法的典型应用。在这些用于解决集体分类的方法中，迭代分类算法(Iterative Classification Algorithm, ICA)[13,14]、Gibbs抽样算法(Gibbs Sampling Algorithm, GSA)[15]以及关系邻居分类器(Relational Neighbor Classifier, RNC)[16]等是3种最具代表性的方法。ICA方法可以利用一个本地分类器迭代地对未分类实例对象进行分类，并且可以综合利用实例对象的属性信息以及从其它关联实例对象获取的关系信息。ICA的每一次迭代步骤都涉及同时更新关联实例的类别标签以及关系特征信息。当所有待分类实例具有稳定的类别标签分布或者迭代次数满足指定的阈值时，整个迭代分类过程结束。本地分类器可以采用多种分类模型进行构建，例如贝叶斯分类模型、逻辑回归模型以及支持向量机模型等。GSA方法也属于迭代分类方法，需要通过基于分类实例属性信息以及关系网络信息训练本地分类器对实例进行分类。在每一次迭代中它具有一个抽样计数过程累计各标签被指派给某个实例的次数，当迭代过程结束时，具有最大累计数值的类别标签即为对应实例最终的分类结果。RNC是一个简单有效的集体分类方法，通过基于直接邻居实例的加权类别标签概率分布对目标实例进行分类，该分类方式是“OSN上两个互相连接的用户之间存在类别相似性”这种源于社会学的同质性原理(homophily)的直接应用。以上3种方法本质上都属于迭代标签传播分类方法，共同的原理是在给定部分实例的类别标签的条件下，利用实例间存在的关系网络信息把类别标签传播到其余未标签实例。文献[16]指出与其它复杂的分类模型相比RNC具有更好的分类性能，应该作为一个基准分类器来评价其它关系网络数据分类模型，因此本文选择RNC作为其中一种基准分类器评价来MLCMRW方法的分类性能。

3 OSN用户分类问题的形式化定义

具有关系网络信息的OSN适合采用图的形式表达，不失一般性，本文将OSN建模为一个加权无向图 $G(V,E,W,X,L,Y)$，其中节点集 $V=\{v_1, v_2, \cdots, v_n\}$ 对应为OSN用户。为方便表述，本文等价交叉使用“节点”和“用户”两个称谓。E 为边的集合，对应OSN用户之间存在的社交关系，并且 $E \subseteq V \times V$。W 为 E 对应的权重矩阵，$\forall w_{ij} \in W$ 表示为节点 v_i 与 v_j 之间边的权重值，W 实质上对应为OSN用户的关系网络特征矩阵。由于OSN用户自身具有的属性特征信息也可以

用于分类，因此对于每一个节点 $v_i \in V$，都分配一个对应的 d 维输入空间的属性特征向量 $x_i=(t_{i1},t_{i2},\cdots,t_{id}) \in \mathbb{R}^d$，其中 t_{ik} 表示为节点 v_i 在第 k 个属性上的取值。$X=[x_1,x_2,\cdots,x_n]^T$ 表示节点的属性特征向量矩阵。假设 $L=\{l_1,l_2,\cdots,l_q\}$ 为类别标签集合，另外考虑到 OSN 用户的分类结果可能有多个类别标签，对每一个节点 $v_i \in V$ 都分配一个标签分布向量 $y_i=(c_{i1},c_{i2},\cdots,c_{iq})$，其中 $c_{ik} \in [0,1]$ 表示分配标签 l_k 给节点 v_i 的概率，y_i 表示分配每一个标签给节点 v_i 的概率集合。矩阵 $Y=[y_1,y_2,\cdots,y_n]^T$ 则表示分配每一个标签给所有节点的概率集合。基于以上定义，OSN 用户分类问题形式上指的是如何利用用户属性特征向量矩阵 X 以及关系网络特征矩阵 W 推断每一个用户节点 $v_i \in V$ 的类别标签分布向量 y_i。

4 OSN 用户分类方法 MLCMRW

4.1 方法概述

MLCMRW 把 V 划分为两个子集 V_l 和 V_u，其中 V_l 中的节点有初始化类别标签而 V_u 中的节点没有初始化类别标签，目标是通过利用节点间存在的联系信息把 V_l 中的标签传播给 V_u 中的节点。假设 V_l 中的节点都为吸收状态，即在任意一个节点 $v_i \in V_l$ 上的随机游走总是在 v_i 上循环，不会跳转到另外一个节点。如果对所有节点进行排序，把有标签的节点排列在无标签节点的前面，那么图 G 的随机游走概率转移矩阵 P 可以重写为如下所示分块矩阵的形式：

$$P=\begin{pmatrix} P_{ll} & P_{lu} \\ P_{ul} & P_{uu} \end{pmatrix}=\begin{pmatrix} I & 0 \\ P_{ul} & P_{uu} \end{pmatrix} \tag{5}$$

其中，P_{ll} 为有标签节点之间的概率转移矩阵区块，因为一次随机游走到达有标签节点时将终止在该节点，所以 P_{ll} 可以等价于单位矩阵 I 的形式；P_{lu} 是一个元素全为 0 的矩阵。由式(5)还可以推出图 G 的 t 步转移概率矩阵 P^t 为：

$$P^t=\begin{pmatrix} I & 0 \\ \sum_{i=1}^{t} P_{uu}^{i-1}P_{ul} & P_{uu}^t \end{pmatrix} \tag{6}$$

当 $t \to \infty$ 时，可获得稳定的转移概率分布 P^∞，且有：

$$P^\infty=\begin{pmatrix} I & 0 \\ (I-P_{uu})^{-1}P_{ul} & P_{uu}^\infty \end{pmatrix} \tag{7}$$

根据第 3 节的问题定义，Y 为记录了所有节点标签分布概率的矩阵，那么可假设 $Y=[Y_l,Y_u]^T$，其中 Y_l 对应 V_l 中所有节点的标签分布矩阵区块，Y_u 对应于 V_u 中所有节点的标签分布矩阵区块。MLCMRW 方法使用 P^∞ 表示节点之间的类别标签相似性，并通过如下公式预测最终节点类别标签分布 $\tilde{Y}$：

$$\tilde{Y}=P^\infty Y=\begin{pmatrix} I & 0 \\ (I-P_{uu})^{-1}P_{ul} & P_{uu}^\infty \end{pmatrix}\begin{pmatrix} Y_l \\ Y_u \end{pmatrix} \tag{8}$$

因为 $P_{uu}^\infty=0$，如果假设 Y_l 是不可以改变的，即已知的初始化类别标签不可更新，可以由式(8)推导得到未知类别标签节点的分类结果 $\tilde{Y}_u$ 为：

$$\tilde{Y}_u=(I-P_{uu})^{-1}P_{ul}Y_l \tag{9}$$

式(9)对一个标签连通图(label-connected)的节点进行分类是有效的，是因为从任何一个无类别标签节点开始的随机游走都可以在有限步内到达任何一个有标签节点。但现实中的 OSN 存在大量孤立子图，如果在这些子图中没有任何具有初始化标签的节点，那么这些子图中的所有节点将不会分配任何标签，即它们最终的标签分布矩阵为空，但事实上 OSN 图上的任何一个节点都可以从它本身具有的属性特征信息中学习得到初始化标签。此外 Y_l 不会变化是不合理的，因为 Y_l 中节点的初始化标签并不能保证完全正确，仍然可以从图的关系网络特征信息中进一步学习提高分类标签的准确性。对于存在的以上两个问题，MLCMRW 方法将分类过程划分为引导和迭代推理两个阶段进行进一步解决。引导阶段从每一个节点的属性信息中学习初始化分类标签；在迭代推理阶段，基于随机游走的迭代推理算法用于更新每一个节点的标签分布。标签迭代更新算法在没有节点的标签分布有变化时或者迭代次数满足指定阈值时可以终止。两个阶段的具体操作过程描述如下：

(1)引导阶段。只使用节点属性特征信息，分配初始化类别标签给每一个 V 中的节点。MLCMRW 采用贝叶斯多项式文本分类模型学习每一个节点的初始化标签分布。

(2)迭代推理阶段。该阶段迭代地应用推断算法对每一个 V 中的节点进行分类，一直到终止条件满足。在步骤 t，每一个节点都采用步骤 $t-1$ 中邻居节点的标签分布加权和作为其在步骤 t 中产生的标签分布。

4.2 引导阶段

引导阶段用于学习用户的初始化标签分布，每一个用户的初始化标签分布可以使用用户属性特征信息学习得到。在 OSN 中，用户属性能从对应用户的 Profile 文档中提取，用户 Profile 文档包含用户的人口统计学信息、教育背景、工作经历、个人兴趣以及偏好等，并且具有一定的真实性。从用户属性集中学习初始化类别标签可以转化为经典的文档分类问题，每一个用户 Profile 文档可以首先基于词典进行分词处理，如用户 i 对应的 Profile 文档 doc_i 可以表示为 $doc_i=\{w_1,w_2,\cdots,w_d\}$，其中 w_j 可以使用词汇 w_j 在 doc_i 中的词频表示。MLCMRW 方法采用多项式贝叶斯文本模型(multinomial naïve bayes，MNB)作为用户 Profile 文档分类器并且通过已标注有多个标签构成的用户 Profile 文档集合进行训练。MNB 是一种快速而又有效的文本分类模型，在 MNB 分类模型中，用户 i 对应的 Profile 文档 doc_i 可以表示为独立于文档长度 $|doc_i|$ 的词项多项式分布，对于给定的任一类别标签 $l_h \in L$，文档 doc_i 属于该标签的概率可以表示为 $p(l_h|doc_i)$，$p(l_h|doc_i)$ 的具体计算方法可以参考文献[17-19]，那么对于待分类的用户 i，其初始化类别标签分布向量 y_i 可以表示为：$y_i=(p(l_1|doc_i),p(l_2|doc_i),\cdots,p(l_q|doc_i))$。

4.3 基于随机游走的分类标签迭代推理

迭代推理阶段首先需要创建权重矩阵 W，对于每一元素 $w_{ij} \in W$，可以通过计算对应用户之间的 Profile 文档相似度作为它的值。根据第 3 节的定义，节点 v_i 的属性向量为 $x_i=(t_{i1},t_{i2},\cdots,t_{id}) \in \mathbb{R}^d$，其中 t_{ik} 可以重新定义为：

$$t_{ik}=f_{ik}\log\left(\frac{n}{idf_k}\right) \tag{10}$$

其中，f_{ik}、idf_k 和 n 分别表示词项 w_k 在节点 v_i 对应文档中的词频、包含词项 w_k 的文档数量以及图 G 的节点数量，那么节点 v_i 和 v_j 对应的 Profile 文档之间的相似度可以使用向量 x_i 和 x_j 的余弦相似度表示，即：

$$\mathrm{sim}(v_i,v_j)=\mathrm{cosine}(x_i,x_j)=\frac{\langle x_i \cdot x_j \rangle}{\|x_i\|\,\|x_j\|} \tag{11}$$

其中，x_i 和 x_j 分别对应节点 v_i 和 v_j 的属性特征向量，$\langle * \cdot * \rangle$ 为两个向量的点积，$\| * \|$ 表示向量的欧几里得范式。w_{ij} 定义如下：

$$w_{ij}=\begin{cases}0, & v_i\neq v_j,(v_i,v_j)\notin E\\ 1, & v_i=v_j\\ \mathrm{sim}(v_i,v_j), & v_i\neq v_j,(v_i,v_j)\in E\end{cases} \tag{12}$$

在 w_{ij} 已知的情况下，应用式(12)可以获得图 G 的概率转移矩阵 P。由于 V_l 的节点需要更新已知分类标签，因此这些节点不再定义为吸收状态，那么根据式(7)计算 P^{∞} 将不可行，从而不能根据式(8)预测节点类别标签。事实上，式(8)可以等价转化为迭代推理预测分类标签分布的形式：

$$Y^t=PY^{t-1} \tag{13}$$

其中，Y^t 表示在图 G 进行 t 步随机游走后的节点标签分布，可以推断节点 v_i 在 t 步随机游走后的标签分布 y_i^t 为：

$$y_i^t=W_iPW^{-1}Y^{t-1} \tag{14}$$

其中，$W_i=(w_{i1},w_{i2},\cdots,w_{in})$ 为 W 的第 i 个行向量。使用式(14)可以在每一次随机游走后集体更新 Y 中的每一标签分布向量，如果标签分布稳定或者随机游走步数达到最大值，则可以停止更新每一个标签分布向量。在实际应用中，基于随机游走的迭代推理算法可以使用准则 $\|Y^t-Y^{t-1}\|\leqslant\varphi$ 作为迭代过程的收敛条件，其中 φ 是预先设定的阈值，如果 Y^t 和 Y^{t-1} 之间的欧几里得范式不大于 φ，那么迭代过程可以终止。MLCMRW 方法的具体过程如算法 1 所示。

算法 1 MLCMRW 算法描述

输入：G(V,E,W,X,L,Y),V_l,V_u,φ

训练：使用 V_l 中的节点属性特征向量信息用于训练 MNB 分类器 MNB_classifier;

引导阶段：//只使用属性特征向量信息学习节点初始化标签分布

for each $v_i\in V$ do

y_i=MNB_classifier(X_i);

迭代推理阶段：//使用关系网络特征信息迭代推理节点标签分布

$Y^0=Y$;t=0;

repeat

t=t+1;

for each $v_i\in V$ do

$y_i^t=W_iPW^{-1}Y^{t-1}$;

until $\|Y^t-Y^{t-1}\|\leqslant\varphi$

$\tilde{Y}=Y^t$;

输出：$\tilde{Y}$

5 实验分析及评价

5.1 数据集

本文选择了 3 个数据集对 MLCMRW 方法进行测试和评价，其中一个是 SCHOLAT 数据集，其余数据集为 DBLP-A 和 DBLP-B，表 1 对这些数据集的特征进行了概括描述，更详细的信息分别描述如下：

(1)SCHOLAT 数据集。该数据集来自于学者网 OSN。学者网提供个人学术信息管理以及主页服务，并且为注册用户提供社交网络服务。为了与其它适用于标签连通图的 OSN 用户分类方法进行比较，本文从学者网全局社交网络图中提取了最大的连通子图作为 SCHOLAT 数据集，该数据集同时包含用户公开的 Profile 信息。数据集中的每一个用户具有一个或多个来自用户自己输入确定的研究兴趣类别标签，所以这些类别标签信息具有很好的准确性。

(2)DBLP-A 和 DBLP-B 数据集。DBLP 数据库提供计算机科学领域的期刊与会议目录信息，本文提取了 DBLP 所包含的一个合著者关系网络，该网络以作者作为节点，并且这些作者都在如下研究领域的代表性学术会议上至少发表 3 篇论文以上，这些会议为：

数据库：ICDE,VLDB,SIGMOD,PODS,EDBT

数据挖掘：KDD,ICDM,SDM,PKDD,PAKDD

人工智能：IJCAI,AAAI

信息检索：SIGIR,ECIR

机器学习：ICML,ECML

算法理论：STOC,FOCS,SODA,COLT

自然语言处理：ACL,ANLP,COLING

分布式并行处理：PODC,ICS

如果某作者在以上任一个会议发表文章，则可以认为该作者对会议所涉及的研究领域感兴趣，所以给其分配相应的研究领域标签，这样每一个作者都可以从以上 8 个研究领域获取至少 1 个研究领域标签。另外，本文还提取了每一个作者发表的文章标题作为对应节点的虚拟 Profile 文档。为了评价 MLCMRW 方法在不同条件下的分类性能，本文从以上 DBLP 合著者关系网络提取了最大的连通子图作为 DBLP-A 数据集用于与其它适用于标签连通图的分类方法进行比较，整个 DBLP 合著者网络作为 DBLP-B 数据集用于验证 MLCMRW 方法是否适用于现实中的无标度 OSN 用户分类。

表 1 实验数据集

特征	数据集		
	SCHOLAT	DBLP-A	DBLP-B
节点数量	432	14183	15036
边数量	860	14957	57531
类别标签数量	36	8	8
平均类别标签数量	1.49	1.91	1.89

5.2 评价准则

OSN 用户分类属于多标签分类问题，需要比传统的单一标签分类模型具有更复杂的性能评价准则，本文采用文献[20]提出的评价准则 Hamming loss、Subset 0/1 Loss、Micro-F1 以及 Macro-F1 评价各分类方法在数据集 SCHOLAT、DBLP_A 与 DBLP_B 上的性能。

5.3 比较方法

为了验证 MLCMRW 的分类性能，本文选取了 AOM 以及 RNC 分类方法在 SCHOLAT 和 DBLP_A 数据集上进行了分类实验结果比较。

(1)Attribute-only method (AOM)：该方法只使用节点的属性特征信息来执行分类，本文采用 MNB 分类模型作为该方法的分类器。

(2)Relational neighbor classifier (RNC)：RNC 是一个有效的关系预测模型，只使用邻居节点的类别标签信息进行预测，不需要学习以及使用本身具有的属性特征信息。在文献[16]中，已证明该方法比其它更复杂的分类模型如概率关系模型以及关系概率树模型等更有效，并被推荐作为基准分类器用于评价其它关系网络数据分类方法的性能。

本文共进行了两种类型的比较实验，一种是分类性能比较实验，考虑到 RW 和 RNC 方法本质上都属于迭代分类方法，为比较公平，这两类方法的迭代次数都统一设置为 10；另外一种是收敛率比较实验，因为 AOM 方法并不属于迭代类型的分类方法，所以只对 MLCMRW 和 RNC 方法进行了比较，并且为方便展示比较结果，只选择了 Micro-F1 作为唯一评价准则。

5.4 实验比较结果和评价

在实验中，对每一个数据集都进行了 5 次交叉验证，并取

平均值作为每一种分类方法的实验结果。性能评价结果以及每一种方法在每一种评价准则的性能排序如表 2 和表 3 所列，收敛率比较结果如图 1 和图 2 所示。在表 2 和表 3 中，符号“↓”表示值越小则对应的性能结果就越好，符号“↑”表示值越大则对应的性能结果就越好。

表 2　SCHOLAT 数据集上的性能比较结果

方法	评价准则				评价排名
	Hamming Loss↓	Subset 0/1 Loss↓	Micro-F1↑	Macro-F1↑	
MLCMRW	0.024 (1)	0.199 (1)	0.753 (1)	0.021 (1)	1
RNC	0.044 (3)	0.394 (2)	0.641 (2)	0.017 (2)	2
AOM	0.042 (2)	0.653 (3)	0.551 (3)	0.015 (3)	3

表 3　DBLP-A 数据集上的性能比较结果

方法	评价准则				评价排名
	Hamming Loss↓	Subset 0/1 Loss↓	Micro-F1↑	Macro-F1↑	
MLCMRW	0.108 (1)	0.206 (1)	0.833 (1)	0.119 (1)	1
RNC	0.122 (2)	0.335 (2)	0.691 (2)	0.103 (2)	2
AOM	0.144 (3)	0.604 (3)	0.653 (3)	0.093 (3)	3

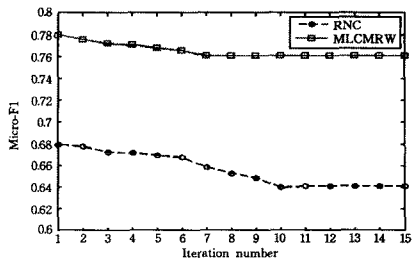

图 1　SCHOLAT 数据集上的收敛率比较

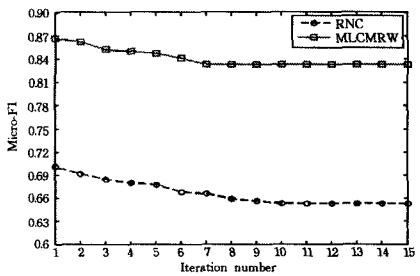

图 2　DBLP-A 数据集上的收敛率比较

从表 2 和表 3 的比较结果可以看出，AOM 方法由于只考虑节点属性特征信息而忽略节点之间的关联信息进行分类，其最终分类性能均比其它方法差，这进一步证明，OSN 中的用户节点并不满足独立一致性假设，并不能直接应用传统的分类方法对其分类。特别地，SCHOLAT 数据集并没有足够的样本训练 AOM 的 MNB 分类模型，所以它在能直接反映分类准确度的两个评价准则 Subset 0/1 Loss 和 Macro-F1 上的实验结果都比较差，但有使用节点间的关系网络信息的 MLCMRW 和 RNC 方法都具有很好的表现。从表 2 和表 3 同样可以看出 MLCMRW 表现比 RNC 好，因为 MLCMRW 同时使用节点的属性信息以及关系网络信息进行分类，该结果也证明同时使用两类信息进行分类比单一使用其中一种信息的分类结果要好。从图 1 和图 2 可以看出，MLCMRW 比 RNC 方法收敛快，因为每一个节点的初始化标签分布可以加快 MLCMRW 方法在迭代推理阶段的收敛速率。另外由于 MLCMRW 在第 7 次迭代时开始收敛而 RNC 在第 10 次迭代时开始收敛，因此在进行实验比较时，两种方法的最大迭代次数都设置为 10 是合理的。

5.5　DBLP-B 数据集的实验结果

表 1 描述的 DBLP-B 数据集具有 15036 个节点和 57531 条边，而且 DBLP-B 对应的关系网络图并不是一个连通图，其包含 326 个连通子图。RNC 方法并不适合直接应用于 DBLP-B 社交网络图的节点分类，因为该方法必须利用节点间存在的关系网络信息传播分类标签，如果节点是孤立的或者连通子图内没有初始化标签，那么 RNC 方法将不具备好的分类结果。而 MLCMRW 方法适合对 DBLP-B 进行节点分类，原因在于 MLCMRW 方法在引导阶段可以分配初始化标签给各节点，然后在迭代推理阶段通过在每一个连通子图中运行迭代推理算法对所有节点进行分类预测，当整体的节点标签分布稳定时，分类过程可终止。本文分别使用 MLCMRW 和 AOM 方法在 DBLP-B 数据集中进行实验，相关实验结果如表 4 所列。可以看出，MLCMRW 方法在 DBLP-B 数据集上仍然具有较好的分类性能，可以用于现实中的无标度 OSN 的用户分类。

表 4　MLCMRW 方法在 DBLP-B 数据集上的性能评价结果

方法	评价准则				评价排名
	Hamming Loss↓	Subset 0/1 Loss↓	Micro-F1↑	Macro-F1↑	
MLCMRW	0.102(1)	0.213(1)	0.812(1)	0.123(1)	1
AOM	0.126(2)	0.504(2)	0.672(2)	0.098 (2)	2

结束语　本文首先对 OSN 的用户分类问题进行了研究，分析了现有代表性方法存在的问题，然后提出一种基于随机游走的分类方法 MLCMRW 对 OSN 用户进行分类。MLCMRW 方法包括学习初始化标签分布以及迭代推理获得稳定标签分布两个阶段，能够综合利用用户属性信息以及关系网络信息进行分类。相关实验结果证明，与其它典型的代表性解决方法对比 MLCMRW 方法更加有效。下一步将研究更多策略以提高 OSN 用户分类性能，例如融合用户交互信息和用户属性信息产生转移概率矩阵，在迭代推理阶段按照某种排序策略对用户节点进行分类从而加快迭代过程，另外，将与更多复杂、高级的分类模型进行比较，以进一步证明 MLCMRW 方法的优势。

参 考 文 献

[1] Arkaitz Z, Christian K, Markus S. Tags vs Shelves: from social tagging to social classification[C]// Proceedings of the 22nd ACM conference on Hypertext and Hypermedia. ACM, 2011: 93-102

[2] Delip R, David Y, Abhishek S, et al. Classifying latent user attributes in twitter[C]// Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents. ACM, 2010: 37-44

[3] Pennacchiotti M, Popescu A-M. A Machine Learning Approach to Twitter User Classification[C]// Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. AAAI, 2011: 281-288

[4] Wu Z. User classification and relationship detecting on social network site Control[C]// Proceedings of 1st International Conference on Automation and Systems Engineering. IEEE, 2011: 1-4

[5] Tang L, Liu H. Leveraging social media networks for classification[J]. Data Mining and Knowledge Discovery, 2011, 23(3): 447-478

[6] Francisco P. A model to classify users of social networks based on pagerank[J]. International Journal of Bifurcation and Chaos,

2012,22(7):1-14
[7] 学者网[EB/OL]. http://www.scholat.com
[8] Shi X X, Li Y, Yu P S. Collective prediction with latent graphs [C]// Proceedings of the 20th ACM Internation conference on Information and knowledge management. ACM, 2011: 1127-1136
[9] Aggarwal C. Social network data analytics[M]. Springer press, Berlin, German, 2011
[10] Rabelo J, Prudencio R B C, Barros F. Collective classification for sentiment analysis in social networks[C]// Proceedings of the 24th International Conference on Tools with Artificial Intelligence. IEEE, 2012, 1: 958-963
[11] Laorden C, Sanz B, Santos I. Collective classification for spam filtering[C]// Proceedings of the 4th International Conference on Computational Intelligence in Security for Information Systems. Springer. 2011: 1-8
[12] Guan J H, Liu H, Xiong W, et al. Effectively predicting protein functions by collective classification-An extended abstract[C]// Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops. IEEE, 2012: 634-639
[13] McDowell L, Gupta K M, Aha D W. Cautious collective classification[J]. Journal of Machine Learning Research, 2009, 10(12): 2777-2836
[14] Neville J, Jensen D. Iterative classification in relational data [C]// Proceeding of the AAAI 2000 Workshop on Statistical Relational Learning of the National Conference on Artificial Intelligence. 2000: 42-49
[15] Kazienko P, Kajdanowicz T. Label-dependent node classification in the network[J]. Neurocomputing, 2012, 75(1): 199-209
[16] Macskassy S A, Provost F J. A simple relational classifier[C]// Proceedings of the 2nd Workshop on Multi-Relational Data Mining. 2003: 64-76
[17] Kibriya A M, Frank E, Pfahringer B, et al. Multinomial naive bayes for text categorization revisited[C]// Proceedings of 17th Australian Joint Conference on Artificial Intelligence. Springer, 2005: 488-499
[18] Su J, Shirab J S, Matwin S. Large scale text classification using semisupervised multinomial naive bayes[C]// Proceedings of the 28th International Conference on Machine Learning. 2011: 97-104
[19] de Campos L M, Fernández-Luna J M, Huete J F, et al. Link-based text classification using bayesian networks[J]. Lecture Notes in Computer Science, 2010, 6203: 397-406
[20] Kong X, Shi X, Philip S Y. Multi-label collective classification [C]// Proceedings of 2011 SIAM International Conference on Data Mining. SDM, 2011: 618-629

(上接第 176 页)

表 2 设计模式识别结果

DesignPatterns	Junit	JHotDraw	JreFactory
FactoryMethod	—	1	2
Proxy	—	—	—
Strategy	1	9	17
AdapterClass	—	—	4
AdapterObject	—	10	34
Observer	—	6	7
Total	1	26	64
Time(ms)	9774	48874	111856

通过对部分设计模式识别结果分析得出，该设计模式识别方法还存在一定的不足：(1)对设计模式识别的准确率有待提高。待识别源代码类结构和设计模式之间的特征匹配如果不够充分就会增加识别结果的错误肯定率；反之，则会增加识别结果的错误否定率。(2)对设计模式异构体或变体的识别考虑不足。每一个设计模式虽然都有一个标准和公认的结构特征，但往往有多种不同的特征表现形式，例如观察者(Observer)模式。这种不同的表现形式可以称为异构体或变体，这对设计模式的识别提出了更高的要求。因此，今后需要对设计模式异构体之间的特征做更细致的分析，从而得到更准确高效的设计模式识别方法。

结束语 本文利用 DPDLXS 语言描述设计模式信息和待抽取的源代码信息，提出源代码间关联度的概念，构建基于关联度的关联类集合，旨在减小设计模式的搜索空间，一定程度上提高识别效率；根据设计模式的特征约束和构建的关联类集合，提出基于关联度和特征约束的设计模式识别算法；最后，将该设计模式识别算法应用于 Junit、JHotDraw 和 JreFactory 3 个开源应用程序的设计模式识别，结果表明该设计模式识别算法具有较高的召回率和效率。

参 考 文 献

[1] Gamma E, Helm R, Johnson R, et al. Design Patterns-Elements of Reusable Object-Oriented Software[M]. New Jersey: Addison-Wesley, 1995
[2] Krämer C, Lutz Prechelt. Design recovery by automated search for structural design patterns in object-oriented software[C]// Proceedings of the Third Working Conference on Reverse Engineering. 1996: 208-215
[3] Rasool G, Mader P. Flexible Design Pattern Detection Based on Feature Types[C]// 26th IEEE/ACM International Conference on Automated Software Engineering (ASE). 2011: 243-252
[4] Balanyi Z, Ferenc R. Mining Design Patterns from C++ Source Code[C]// Proc. Internationl Conf. on Software Maintenance (ICSM'03). 2003: 305-314
[5] Dobis M, Majtas L. Mining Design Patterns from Existing Projects Using Static and Run-Time Analysis[J]. Springer Software Engineering Techniques, 2011, 4980: 62-75
[6] 古辉，张炜星. 基于 XML Schema 技术的设计模式定义方法[J]. 计算机科学，2014，41(1)：254-257
[7] Antoniol G, Fiutem R, Cristoforetti L. Design pattern recovery in object-oriented software[C]// Proceedings of 6th IEEE International Workshop on Program Comprehension (IWPC 1998). 1998: 153-160
[8] Gueheneuc Y G, Guyomarc'h J Y, Sahraoui H. Improving design-pattern identification: a new approach and an exploratory study[J]. Software Quality Journal, 2010, 18(1): 145-174
[9] 苗康，余啸，赵吉，等. 基于关系演算的 Java 模式识别[J]. 计算机应用研究，2010，27(9)：3425-3430
[10] Li F, Li Q S, Su Y, et al. Detection of design patterns by combining static and dynamic analyses[J]. Journal of Shanghai University(English Edition), 2007, 11(2): 156-162
[11] 严蔚敏，吴伟民. 数据结构(c 语言版)[M]. 北京：清华大学出版社，1997
[12] 阎宏. Java 与模式[M]. 北京：电子工业出版社，2002