基于链路预测的微博用户关系分析

傅颖斌 陈羽中
（福州大学福建省网络计算与智能信息处理重点实验室 福州 350108）

摘 要 随着以微博为代表的在线社交网站的发展，微博用户之间形成了复杂的社会网络。针对微博社会网络，研究了影响微博用户之间关系形成的各种因素，提出了基于链路预测的微博用户关系分析模型。首先分析了网络结构特征在微博社会网络中的作用，同时针对微博社会网络的特点，引入微博属性特征，构造基于随机森林的链路预测模型，并将模型应用于新浪微博用户数据集，进行微博用户关系的训练预测，通过比较引入微博属性特征前后的预测性能以及特征的重要性分布，分析了各类特征对微博用户关系形成的影响，揭示了除传统的网络结构特征外，微博属性特征对微博用户关系的形成具有重要的影响力。

关键词 链路预测，社会网络，微博属性，随机森林
中图法分类号 TP393 **文献标识码** A

Relationship Analysis of Microblogging User with Link Prediction

FU Ying-bin CHEN Yu-zhong
(Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, Fuzhou University, Fuzhou 350108, China)

Abstract With the development of online social networking sites represented by microblog, the microblogging users form some complex social networks. In order to study the factors that affect the formation of relationship among microblogging users, this paper used link prediction to analyze the relationship of micro-blogging users. Firstly, this paper studied how the features of network structure affect the formation of microblogging network. The features of microblogging attribute were also analyzed and introduced to build a link prediction model based on random forest classifier. The link prediction model was tested on a user data set collected from Sina Weibo. By comparing the prediction performance with and without the introduction of microblogging attribute features and analyzing the importance distribution of features, we found that besides the network structure features, microblogging attribute features have significant effect on the formation of user relationship, and can improve the prediction performance significantly.

Keywords Link prediction, Social network, Microblogging attribute, Random forest

1 引言

随着移动互联网技术和 Web 技术的发展，以微博为代表的在线社交网站成为了人们日常交流、娱乐、通信的重要工具。全球最早的微博平台 Twitter，经过短短两三年的发展，成为了最大的社交平台，用户数超过 5 亿。而在中国，根据 DCCI 互联网数据中心发布的《2012 中国微博蓝皮书》显示，中国微博用户数已经达到 3.27 亿。与传统社会网络服务（Social Network Services，SNS）不同的是，微博用户可以随意关注任何其他用户，这使得微博用户之间构成一个结构复杂、联系紧密的社会网络。微博用户关系作为构建微博社会网络的基础，极大影响了微博社会网络的形成和发展，因此，微博用户关系的分析对微博社会网络的研究具有十分重要的理论价值和实践意义。

网络中的链路预测是指如何通过已知的网络结构等信息预测网络中尚未产生连接的两个节点之间产生连接的可能性[1]。在线社交网站中用户和用户关系构建出庞大的社会网络，网络中的顶点代表用户，边代表用户关系，链路预测问题正是对用户未来关系的分析。Liben-Nowell 和 Kleinberg[2] 最早提出了应用于社会网络的链路预测模型，其通过计算基于图的社会网络模型中顶点间的拓扑结构相似性来预测链路形成的可能性，同时对这些拓扑结构特征进行分析。目前，社会网络链路预测模型主要发展为三大类：

1）基于有监督学习的分类模型，如决策树、朴素贝叶斯、神经网络、SVM、KNN 以及集成方法中的 bagging、boosting 和随机森林等。这类模型是链路预测问题的主流方法，Hasan 等比较了不同机器学习算法对学术论文数据集（DBLP）的链路预测效果[3]；Michael 等人则将分类模型应用于社会网络的链路预测问题[4]，文中分析出一些简单、易于计算的特征集合，特别提出朋友评价（Friends-measure）特征，并

到稿日期：2013-04-11 返修日期：2013-06-24 本文受福建省自然科学基金（2013J01232），福建省教育厅重点项目（JK2012003），福建省科技创新平台项目（2009J1007）资助。

傅颖斌（1988－），男，硕士生，主要研究方向为社会网络分析、数据挖掘，E-mail：fuyingbin@outlook.com；陈羽中（1979－），男，博士，副教授，主要研究方向为计算智能、数据挖掘、复杂网络。

用这些特征对 Facebook、Academia、Flickr、TheMarker、YouTube 的大规模数据集进行训练预测，比较不同分类算法的预测结果。

2）概率模型，该模型主要是建立一组可调参数的模型，然后使用优化策略寻找最优的参数值，使模型能够达到最优，这时两个未连边的节点对的概率就是它们产生连边的条件概率。概率模型的构建方法有贝叶斯网络模型[5]和马尔科夫网络[6]关系模型等。Hopcroft 和清华大学的唐杰等人[7]提出了基于因子图模型的三元组因子图（Triad Factor Graph）模型，该模型针对 Twitter 中的双向关系（two-way relationship）的预测问题，得到了很好的预测结果。

3）线性代数方法，该方法是通过降阶相似矩阵来计算网络中节点之间的相似性，Kunegis 等人[8]利用图的邻接矩阵，并定义一个函数 F 使得两个时刻的邻接矩阵的差异性最小，这样就将链路预测问题转换成线性代数优化问题，之后再通过矩阵变换和降维的方法将问题转换为一维的最小二乘曲线拟合问题。

链路预测问题作为社会网络研究中的重要研究方向，对于研究和分析社团的演化、社会网络关系的形成有重要帮助，而对于国内目前发展最快速的微博社会网络的研究大都集中在微博社会网络拓扑结构[9]和微博内容挖掘上[10]，对影响微博用户关系形成的特征的研究却比较少。本文运用基于有监督学习链路预测方法，结合网络结构特征和微博属性特征，首先分析了度特征、共同朋友特征等网络拓扑结构特征以及朋友评价（Friends-measure）、邻居子图等特征对用户关系的影响；并分析了用户的关注数、粉丝数、微博消息数、所在地等微博属性特征对微博用户关系的影响，构造出基于随机森林（Random Forest）的链路预测模型，利用所采集的新浪微博用户数据集训练预测，获得了理想的预测结果；最后分析了预测模型中各特征的 Gini 指标，获得了特征的重要性分布，从而验证了网络拓扑结构特征和微博属性特征对用户关系的影响。

本文的主要贡献在于：1）运用链路预测方法对微博社会网络用户关系进行分析；2）在传统链路预测拓扑结构特征的基础上，引入微博属性特征，提高了链路预测的预测性能；3）利用预测模型分析了网络拓扑结构特征和微博属性特征的重要性分布。

2 问题描述

将微博社会网络的拓扑结构看作是图，图中的顶点和边分别代表微博社会网络中的用户和用户关系，与 Facebook 或人人网的网络结构不同，微博社会网络结构是一个有向图结构，因此定义有向图 $G=(V,E)$，其中 V 为顶点集合，E 为边集合。假设顶点 $u,v\in V$，若 $(u,v)\in E$，则表示微博社会网络中存在用户关系：微博用户 u 关注微博用户 v，那么对用户关系的预测问题也就是对图的链路预测问题。对于链路预测问题，本文构建出有监督的分类学习模型，选取训练数据集对分类模型进行训练。对于 $u,v\in V$，设顶点对 (u,v) 的分类标记为 $y^{(u,v)}$，则 $y^{(u,v)}$ 的定义为：

$$y^{(u,v)}=\begin{cases}1, & \text{if } (u,v)\in E\\ 0, & \text{if } (u,v)\notin E\end{cases} \tag{1}$$

可见，该分类学习模型为二元分类模型。从数据集构建的图中抽取顶点对构建样本集合，其中 $y^{(u,v)}$ 值为 1 的顶点对集合构成正样本，值为 0 的顶点对集合构建负样本。再从样本集合中抽取训练集合和测试集合，训练集合对二元分类模型进行训练，测试集合对模型进行验证。由此构建的模型可预测顶点对 (u,v) 是否存在边，即是否存在用户关系。

3 特征集合构造

分类模型的关键在于特征集合的构造，本节主要介绍分类模型中使用的特征。根据特征的来源，特征集合可分成网络结构特征和微博属性特征两类。网络结构特征是指从微博社会网络生成的有向图中的拓扑结构提取的特征，依据文献[1,4]提出的拓扑结构特征，并针对微博社会网络的特点对共同朋友特征、朋友总数特征、Adamic-Adar 特征进行扩展。网络结构特征的计算复杂度较高，并且微博社会网络具有丰富的属性信息，仅网络结构特征不足以反映微博用户之间的关系，因此本文引入微博属性特征，研究了从抓取数据中获得的用户粉丝数、关注数、所在地、微博消息数等微博属性特征对关系形成的影响。分类模型所采用的网络结构特征和微博属性特征介绍如下。

3.1 网络结构特征

顶点度特征：在网络拓扑结构中，顶点度是一个非常重要的衡量指标。对于有向图 $G=(V,E)$，令顶点 $v\in V$，则顶点 v 的邻居顶点集合（$\Gamma(v)$）、链入邻居顶点集合（$\Gamma_{in}(v)$）、链出邻居顶点集合（$\Gamma_{out}(v)$），以及既是链入又是链出的邻居顶点集合（$\Gamma_{bi}(v)$）的定义如下：

$$\begin{aligned}&\Gamma(v)=\{u\mid(u,v)\in E \text{ or } (v,u)\in E\}\\&\Gamma_{in}(v)=\{u\mid(u,v)\in E\}\\&\Gamma_{out}(v)=\{u\mid(v,u)\in E\}\\&\Gamma_{bi}(v)=\Gamma_{in}(v)\bigcap\Gamma_{out}(v)\end{aligned} \tag{2}$$

基于以上不同邻居顶点集合的定义，根据集合中元素的数量可得出顶点 v 的 4 个顶点度特征，分别定义为：

$$\begin{aligned}&d(v)=|\Gamma(v)|\\&d_{in}(v)=|\Gamma_{in}(v)|\\&d_{out}(v)=|\Gamma_{out}(v)|\\&d_{bi}(v)=|\Gamma_{in}(v)\bigcap\Gamma_{out}(v)|\end{aligned} \tag{3}$$

共同朋友特征：也称为共同邻居（CN）特征[1]，是一个重要的链路预测特征。在有向图中，对于顶点对 (u,v)，可以定义 3 个共同朋友特征，分别是：

$$\begin{aligned}&common\_friends_{in}(u,v)=|\Gamma_{in}(u)\bigcap\Gamma_{in}(v)|\\&common\_friends_{out}(u,v)=|\Gamma_{out}(u)\bigcap\Gamma_{out}(v)|\\&common\_friends_{bi}(u,v)=|\Gamma_{bi}(u)\bigcap\Gamma_{bi}(v)|\end{aligned} \tag{4}$$

朋友总数特征：共同朋友特征是两个顶点的邻居顶点的交集大小，朋友总数则是邻居顶点并集大小，与共同朋友特征类似，朋友总数特征也有 3 个，分别是：

$$\begin{aligned}&total\_friends_{in}(u,v)=|\Gamma_{in}(u)\bigcup\Gamma_{in}(v)|\\&total\_friends_{out}(u,v)=|\Gamma_{out}(u)\bigcup\Gamma_{out}(v)|\\&total\_friends_{bi}(u,v)=|\Gamma_{bi}(u)\bigcup\Gamma_{bi}(v)|\end{aligned} \tag{5}$$

中介朋友特征：在社会网络分析（Social Network Analysis，SNA）中，介数中心度（betweenness centrality）是一个很重要的指标，它反映了一个顶点在网络中处于许多路径之上，

具有能够控制其他两个用户交往的能力。同样，若顶点 u 到顶点 v 之间存在着许多中介顶点，那么用户 u 关注用户 v 的可能性也就越大。因此定义中介朋友特征为：

$$transitive\_friends(u,v)=|\Gamma_{out}(u)\bigcap\Gamma_{in}(v)| \tag{6}$$

优先链接特征：在社会网络中存在这样的现象，朋友越多的个体之间变成朋友的可能性越大，也就是“精英”之间形成关系的可能性越大，这可以被称为富者愈富现象，优先链接特征值较大的顶点对形成边的可能性较大，在这里用两个顶点的邻居顶点数的乘积表示：

$$preferential\_attachment\_score(u,v)=|\Gamma(u)|\cdot|\Gamma(v)| \tag{7}$$

Adamic-Adar 特征：Adamic-Adar(AA)指标[11]的思想是度小的共同邻居顶点的贡献大于度大的共同邻居顶点。因而为共同邻居顶点的度赋予一个权重值，该权重值等于该顶点度的对数分之一。对于有向图，这里增加一个共同中介顶点的 Adamic-Adar 特征。因此，该特征的定义如下：

$$\begin{aligned}adamic\_adar(u,v)&=\sum_{z\in\Gamma(u)\cap\Gamma(v)}\frac{1}{\lg d(z)}\\adamic\_adar\_out\_in(u,v)&=\sum_{z\in\Gamma_{out}(u)\cap\Gamma_{in}(v)}\frac{1}{\lg d(z)}\end{aligned} \tag{8}$$

朋友评价(Friends-measure)特征：Michael 等人基于 Katz 指标[12]的思想提出了 Friends-measure 特征[4]。该特征基于这样的假设，若两个顶点的邻居顶点之间存在越多连接，那么这两个顶点之间形成链路的机会也就越高。因此，朋友评价也可以看成是两个邻居顶点间边的数量。朋友评价的定义为：

$$friends\_measure(u,v)=\sum_{x\in\Gamma(u)}\sum_{y\in\Gamma(v)}\delta(x,y) \tag{9}$$

其中 $\delta(x,y)$ 定义为：

$$\delta(x,y)=\begin{cases}1, & \text{if } x=y \text{ or } (x,y)\in E \text{ or } (y,x)\in E\\0, & \text{otherwise}\end{cases} \tag{10}$$

从定义可以看出，朋友评价特征是 Katz 指标当 $\beta=1$ 且 $l_{\max}=2$ 时的特例，但在运算复杂度上却比 Katz 指标小很多。

反向关系特征：在熟人社交圈子中，若某个用户关注你，那么你很可能也会去关注他。也就是说，若 $(u,v)\in E$，那么很可能存在 $(v,u)\in E$，因此该特征是一个二值特征。反向关系特征定义为：

$$oppsite\_direction\_friends(u,v)=\begin{cases}1, & \text{if } (u,v)\in E\\0, & \text{otherwise}\end{cases} \tag{11}$$

邻居子图特征：邻居子图是指由两个顶点的邻居顶点构成的子图，在邻居子图中，若边的数量越大，就可以认为邻居顶点之间的关系越紧密。邻居子图特征就是两个顶点邻居子图中边的数量，定义为：

$$\begin{aligned}&subgraph\_edge\_count(u,v)=\\&\quad|\{(x,y)\in E|x,y\in\Gamma(u)\bigcup\Gamma(v)\}|\\&subgraph\_add\_edge\_count(u,v)=\\&\quad|\{(x,y)\in E|x,y\in\Gamma(u)\bigcup\Gamma(v)\bigcup\{u,v\}\}|\end{aligned} \tag{12}$$

3.2 微博属性特征

网络结构特征从拓扑结构的角度影响用户关系，但微博社会网络中丰富的属性信息对用户关系也具有直接的影响力。微博属性特征是指微博用户的个人属性，微博用户个人属性包括：用户 ID、用户名称、性别、描述、所在地、关注数、粉丝数、消息数等，本文所用的爬盟数据集(见表 1)中包含上述这些属性信息。对于微博用户 u，定义其关注数、粉丝数、消息数分别为：$attention\_num(u)$，$fans\_num(u)$，$message\_num(u)$。对于样本集中的顶点对 (u,v)，令 u 为起始点，v 为终止点，通过顶点的编号就可从数据集中获取相应的属性信息，并绘制出微博属性特征在正负样本中的分布图，限于篇幅，每一类特征选取一个特征绘制特征分布图。微博属性特征可以归纳为以下几类。

关注数特征：关注数量的大小反映了微博用户的活跃程度，相对来说，活跃用户之间形成关系的可能性就大。从图 1 中正负样本起始点的关注数分布可以看出，用户关注数大都集中在 0 到 500 之间，在 2000 处的高峰是由于新浪微博限定用户的关注数上限为 2000。图中也可以看出，负样本中关注数大多集中在 100 以内，而正样本的关注数分布相对负样本较平缓。对于顶点对 (u,v)，关注数特征包括起始点关注数($attention\_num\_s$)、终止点关注数($attention\_num\_d$)、关注数乘积($attention\_num\_bi$)，定义为：

$$\begin{aligned}&attention\_num\_s=attention\_num(u)\\&attention\_num\_d=attention\_num(v)\\&attention\_num\_bi=attention\_num\_s*attention\_num\_d\end{aligned} \tag{13}$$

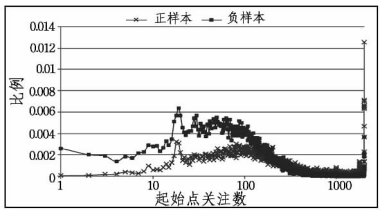

图 1 起始点关注数特征分布图

粉丝数特征：粉丝数特征一方面体现了用户的活跃程度，同时也反映了用户的影响力和知名度，粉丝数越大，其影响力和知名度也越大。图 2 显示了终止点的粉丝数分布，可见负样本中的相对集中，而正样本粉丝数分布相对较平缓，可见粉丝数少，其形成关系的可能性相对较小。对于顶点对 (u,v)，粉丝数特征包括起始点粉丝数(fan_num_s)、终止点粉丝数(fan_num_d)、粉丝数乘积(fan_num_bi)，定义为：

$$\begin{aligned}&fans\_num\_s=fans\_num(u)\\&fans\_num\_d=fans\_num(v)\\&fans\_num\_bi=fans\_num\_s*fans\_num\_d\end{aligned} \tag{14}$$

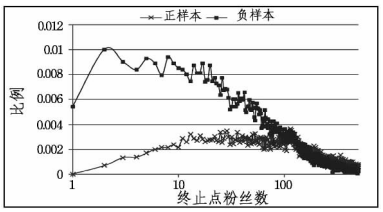

图 2 终止点粉丝数特征分布

消息数特征：微博用户的消息数，也在一定程度上反映了用户的活跃度，活跃用户之间形成关系的可能性相对较大。从图 3 中终止点的消息数分布可以看出，正样本的分布相比负样本更加平缓。对于顶点对 (u,v)，消息数特征包括起始点消息数($message\_num\_s$)、终止点消息数($message\_num\_d$)，定义为：

$$message\_num\_s = message\_num(u)$$
$$message\_num\_d = message\_num(v) \quad (15)$$

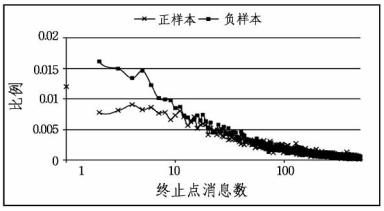

图 3　终止点消息数特征分布

地理位置特征：除了微博用户的关注数、粉丝数、消息数之外，用户的地理位置信息对用户关系的形成也有一定影响，这里的地理位置是指用户设置的所在地信息。在新浪微博中，用户所在地信息主要包括两个等级，第一级(L1)是省级行政区，第二级(L2)是市级行政区，如“福建 福州”。因此，设用户 u 第一级所在地为 $adr(u,1)$，第二级所在地为 $adr(u,2)$，则地理位置特征(address_score)定义如下：

$$address\_score(u,v) = \begin{cases} 2, & \text{if } adr(u,1)==adr(v,1) \text{ and } adr(u,2)==adr(v,2) \\ 1, & \text{else if } adr(u,1)==adr(v,1) \\ 0, & \text{otherwise} \end{cases} \quad (16)$$

对样本集中正负样本的地理位置特征值分布进行统计，结果如图 4 所示。可见，负样本集合中大多数顶点对的所在地特征值都为 0，而所在地特征值为 1 或 2 的边中，正样本比例明显大于负样本，由此可见，所在地特征对微博社会网络中链路的形成有一定关系。

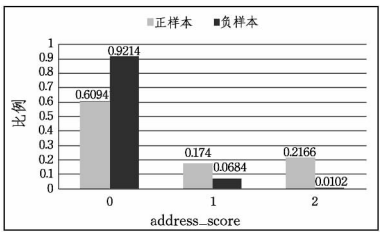

图 4　所在地特征分布

4　算法描述

本实验分类学习算法采用随机森林算法(Random Forests，RF)[13]。RF 是 Breiman 在 2001 年提出的一种组合分类和回归算法，它首先以分类回归树(classification and regression trees，CART)[14]作为元分类器，采用 Bagging(bootstrap aggregation)方法[15]制造有差异的训练样本集，并在构建单棵决策树时采用一种随机子空间划分的策略，从随机选择的部分属性中挑选最佳属性对内部节点进行属性分裂。这种“双随机”的策略使得 RF 中的子分类器之间具有较大的差异性，从而具有优越的分类性能。

在本文中，预测模型为一组元分类器(CART 决策树)构成的随机森林，记作：$\{h(x,\theta_k)\}$。其中 h 表示元分类器，$\{\theta_k\}$ 是相互独立且同分布的随机向量，x 表示输入向量，即本文中每条边所对应的特征向量。随机森林通过 Bagging 方法构建 k 棵决策树，每棵决策树对输入向量进行分类投票，最终确定分类结果，如图 5 所示。

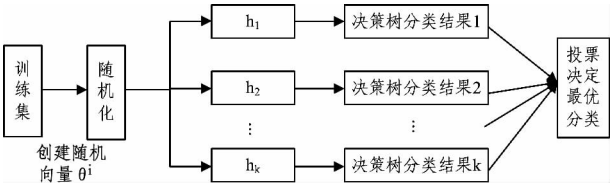

图 5　随机森林示意图

RF 通过构造不同的训练集增加分类模型间的差异性，从而提高组合分类模型的预测能力。最终的分类决策模型为：

$$H(x) = \arg\max_{Y} \sum_{i=1}^{k} I(h_i(x)=Y) \quad (17)$$

式中，H 表示集合分类模型，h 是单个决策树，Y 是输出变量，I 为示性函数[13]。对于传统链路预测模型，输入变量中只包含网络的拓扑结构特征，而通过本文分析，微博属性特征对于分类决策也具有重要的影响，因而在本文特征集合中加入了微博属性特征子集以提高分类预测性能。

对于社会网络拓扑图，其可能存在的边的数量级是顶点的平方，但实际存在的边的数量远小于不存在的边的数量，这就导致了数据偏斜问题，在这种情况下，样本分布不均匀，无法准确反映整个空间的数据分布，分类器容易被大类淹没而忽略小类[16]。对于本实验的二元分类模型，由于社会网络中节点数量往往比较庞大，通常达到十万、百万甚至千万的数量级，其可能存在边的数量级是节点数量的平方，但实际上存在的边的数量占所有边的数量比例很小，由此造成正负样本分布不均匀，正样本数量只占所有样本数量的很小部分，导致数据偏斜问题更加明显[17]。解决数据偏斜问题的主要对策之一就是重采样(re-sampling)，重采样方式主要通过减轻数据集的不均衡程度来平衡正负样本的分布比例，提高少数类的分类性能。重采样方法包括过采样(over-sampling)和欠采样(under-sampling)，过采样通过增加少数类的样本来提高少数类的分类性能，欠采样通过减少多数类样本来提高少数类样本的分类性能。在实验中，考虑到特征计算的复杂性，重采样时对存在的边和不存在的边都采用有放回抽样策略，使得抽取出的样本中正负样本比例平衡(1∶1)。

5　实验步骤与参数

本文采用十折交叉验证法对算法性能进行验证，并通过计算 AUC(area under the ROC curve)指标来判断分类效果，其计算方式为：每次随机从测试集中选取一条边与随机选择的不存在的边进行比较，如果测试集中的边的分数值大于不存在的边的分数值，就加 1 分；如果两个分数值相等，就加 0.5 分。独立地比较 n 次，如果有 n' 次测试集中的边的分数值大于不存在的边的分数，有 n'' 次两分数值相等，则 AUC 定义为：

$$AUC = \frac{n' + 0.5n''}{n} \quad (18)$$

整体实验步骤为：

1. 将社会网络原始数据集构建出边集。

2. 根据边集大小从中抽取正负实验样本，根据本文所用数据集关系数量，取正负样本数都为 25000，即存在的边和不存在的边各随机抽取 25000 条。

3. 对样本进行十折交叉验证，产生 10 组训练集和测试集。

4. 对于每组训练集和测试集，计算特征值，用训练集对随机森林分类器进行训练，再对测试集进行验证，计算 AUC 值。

5. 计算 10 组 AUC 值的平均值。

本文选用中国爬盟[18]的新浪微博用户关系数据集作为微博研究数据集，中国爬盟是通过众包方式获取微博数据的合作组织，用户通过贡献并分享获取的部分数据来换取更多的数据，最终达到共赢的目的。本文选取爬盟 2012 年 6 月 8 日采集的整合用户和关系数据集，对数据集进行清洗，去除其中重复和错误数据，通过关注用户(follow_userid)字段获取用户关系，构建出社会网络拓扑结构。同时，本文还选取文献[4]提供的 3 个国外社会网络数据集，这些社会网络与微博社会网络类似，也能构成有向图结构。数据集的基本信息如表 1 所列。

表 1　数据集

网络	用户数量	关系数量	时间
新浪微博	324,725	675,775	2012-06-08
Academia	200,169	1,398,063	2011
Flickr	1,133,547	7,237,983	2010
YouTube	1,138,499	4,945,382	2007

对于微博数据集，除了网络结构特征集合之外，另加入微博属性特征集合，因此，对于微博数据集，有两个特征集合进行对比，如表 2 所列。

表 2　特征集合

特征名称	特征数
网络结构特征集合	22 个
所有特征集合	31 个(网络结构特征＋微博属性特征)

实验中的 RF 算法采用 scikit-learn 机器学习程序包中的随机森林算法实现，该实现基于 Breiman 随机森林理论，可以设置树的个数和分裂指标。本文选择 Gini 指标作为树的分裂指标，并计算随机森林中各个决策树特征的 Gini 指标的平均值，将其作为特征的重要性指标，从而比较出各个特征对于分类的重要性程度。

6　实验结果与分析

各个数据集的 AUC 结果如图 6 所示，其中 Academia、Flickr、Youtube 的结果与文献[4]结果基本一致；pm0608_1 和 pm0608_2 都是爬盟新浪微博数据集的 AUC 结果，其中，pm0608_1 采用网络结构特征集合，pm0608_2 的特征集合包括网络结构特征和微博属性特征。

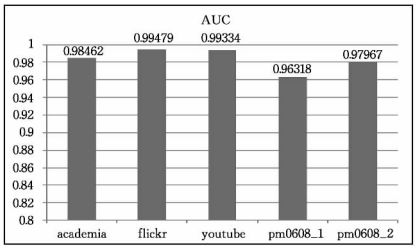

图 6　AUC 结果比较

从 AUC 结果可以看出，只采用网络结构特征，链路预测结果已经能达到较好的效果，而加入微博属性特征后，预测效果有所提高，可见微博属性特征对微博用户关系的形成具有影响力。而从单个特征来看，特征之间的重要性也是不同的。

图 7 列出了各个特征的 Gini 指标分布，其中微博属性有添加标注。从图中可以看出，朋友评价特征、反向关系特征、优先链接特征、邻居子图特征对微博用户关系形成与否影响最大，这与文献[4]中结论相似。而微博属性的 Gini 指标虽然不如上述几个特征，但总体上对微博用户关系形成还是具有较明显的影响力，特别是关注数特征中的起始点关注数、用户所在地特征、粉丝数特征中的终止点粉丝数的影响力较大，此外，其他微博属性如消息数特征也有一定的影响力。粉丝数、关注数、消息数反映出微博用户自身的活跃程度和在微博社会网络中的影响力，越活跃、影响力越大的用户越容易被人关注；而所在地特征反映出地理位置信息对用户关系的影响，在同一地区的用户之间形成关系的可能性越高。

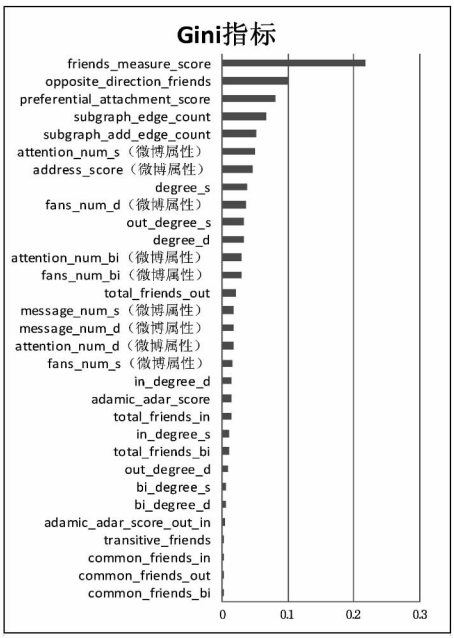

图 7　pm0608_2 特征集 gini 指标分布

结束语　本文采用链路预测的方法对影响微博用户关系的因素进行分析，并根据微博社会网络的特点，着重从网络结构特征和微博属性特征两个方面，构建出基于随机森林分类器的链路预测模型。本文的主要贡献在于：首先将传统链路预测模型中的拓扑结构特征运用于真实微博社会网络中，获得了很好的预测效果，验证了网络结构特征对用户关系形成具有影响力；其次，本文还引入微博属性特征，通过将只有网络结构特征集合的预测结果与包含微博属性特征的所有特征集合的预测结果进行比较，发现加入微博属性特征有助于提高预测性能；最后，比较各个特征的 Gini 指标，得到特征的重要性分布，验证了微博属性特征对用户关系的影响。上述对微博社会网络中影响微博用户关系的特征的分析，对了解微博用户的行为规律和微博信息的传播机制有着重要的意义。

下一步的研究工作主要包括两个方面：在特征方面，增加微博文本内容对用户关系形成的影响力分析；在时间维度上，分析一个时间序列上用户关系的变化，研究微博用户关系的动态演化机制。

参 考 文 献

[1]　吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报，2010，39(5)：652

(下转第 244 页)

扰动；(ii)融合信息计算准确；(iii)实数值融合算法性能优越。本文重点研究 LAC 的聚类融合算法，但是本文所提出的软子空间聚类融合算法的框架不限于 LAC 算法。未来，将深入研究其它软子空间聚类融合算法及其参数之间的关系。另外，本文实验中选取特征子集的比例 $1/por$ 为随机值，对于不同数据库该值可能有所不同，未来我们将同时深入研究不同数据库与选取特征子集的比例 $1/por$ 之间的关系。

参 考 文 献

[1] Kriegel H P, Kroger P, Zimek A. Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering [J]. ACM Transactions on Knowledge Discovery from Data, 2009, 3(1): 1-58

[2] Parsons L, Haque E, Liu H. Subspace Clustering for High Dimensional Data: A Review [J]. ACM SIGKDD Explorations Newsletter-Special issue on learning from imbalanaced datasets, 2004, 6(1): 90-105

[3] Huang J Z. Automated variable weighting in K-means type clustering [J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668

[4] Gan G, Wu J. A convergence theorem for the fuzzy subspace clustering (FSC) algorithm [J]. Pattern Recognition, 2008, 41(6): 1939-1947

[5] Domeniconi C. Locally adaptive metrics fore clustering high dimensional data [J]. Data mining knowledge discovery, 2007, 14: 63-97

[6] Jing L P. An entropy weighting K-means algorithm for subspace clustering of high dimensional sparse data [J]. IEEE Trans. on Knowledge and Data Engineering, 2007, 19(8): 1026-1041

[7] Deng Z H, Choi K S, Chung F L, et al. Enhanced soft subspace clustering integrating within-cluster and between-cluster information [J]. Pattern Recognition, 2010, 43(3): 767-781

[8] Chen X J, Ye Y M, Xu X F, et al. A feature group weighting method for subspace clustering of high-dimensional data [J]. Pattern Recognition, 2012, 45(1): 434-446

[9] Xia Hu, Zhuang Jian, Yu De-hong. Novel Soft Subspace Clustering with Multi-objective Evolutionary Approach for High-dimen-sional Data [J]. Pattern Recognition, 2013, 46(9): 2562-2575

[10] 陈黎飞，郭躬德，姜青山. 自适应的软子空间聚类算法 [J]. 软件学报，2010，21(10)：2513-2523

[11] Domeniconi C, Al-Razgan M. Weighted Cluster Ensembles: Methods and Analysis [J]. ACM Trans. Knowledge Discovery from Data, 2009, 2(4): 1-40

[12] Fred A L N, Jain A K. Combining Multiple Clusterings Using Evidence Accumulation [J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27(6): 835-850

[13] Kuncheva L I, Vetrov D P. Evaluation of stability of k-means cluster ensembles with respect to random initialization [J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2006, 28(11): 1798-1808

[14] Fern Xiaoli Z, Brodley Carla E. Random projection for high dimensional data clustering: a cluster ensemble approach [C] // Proceedings of 20th International Conference on Machine learning (ICML2003). Washington, DC, USA: AAAI Press, 2003: 186-193

[15] Likas A, Vlassis N, Verbeek J J. The Global k-Means Clustering Algorithm [J]. Pattern Recognition, 2003, 36: 451-461

[16] Strehl A, Ghosh J. Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions [J]. Journal of Machine Learning Research, 2002, 3: 583-617

[17] Iam-On N, Boongoen T, Garrett S, et al. A Link-Based Approach to the Cluster Ensemble Problem [J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2011, 33(12): 2396-2409

(上接第 205 页)

[2] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks[J]. Journal of the American society for information science and technology, 2007, 58(7): 1019-1031

[3] Al Hasan M, Chaoji V, Salem S, et al. Link prediction using supervised learning[C] // SDM'06: Workshop on Link Analysis, Counter-terrorism and Security. 2006

[4] Fire M, Tenenboim L, Lesser O, et al. Link prediction in social networks using computationally efficient topological features[C] // Privacy, security, risk and trust (PASSAT), 2011 IEEE third international conference on and 2011 IEEE third international confernece on social computing (SocialCom). 2011: 73-80

[5] Heckerman D, Geiger D, Chickering D M. Learning Bayesian networks: The combination of knowledge and statistical data [J]. Machine learning, 1995, 20(3): 197-243

[6] Taskar B, Abbeel P, Koller D. Discriminative probabilistic models for relational data[C] // Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence. 2002: 485-492

[7] Hopcroft J, Lou T, Tang J. Who will follow you back?: reciprocal relationship prediction[C] // Proceedings of the 20th ACM international conference on Information and knowledge management. 2011: 1137-1146

[8] Kunegis J, Lommatzsch A. Learning spectral graph transformations for link prediction[C] // Proceedings of the 26th Annual International Conference on Machine Learning. 2009: 561-568

[9] 樊鹏翼，王晖，姜志宏，等. 微博网络测量研究[J]. 计算机研究与发展，2012，49(4)：691-699

[10] 周刚，邹鸿程，熊小兵，等. MB-SinglePass：基于组合相似度的微博话题检测[J]. 计算机科学，2012，39(10)：198-202

[11] Adamic L A, Adar E. Friends and neighbors on the web[J]. Social networks, 2003, 25(3): 211-230

[12] Katz L. A new status index derived from sociometric analysis [J]. Psychometrika, 1953, 18(1): 39-43

[13] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32

[14] Olshen R, Breiman L, Friedman J H, et al. Classification and Regression Trees[M]. Wadsworth International Group, 1984

[15] Breiman L. Bagging predictors[J]. Machine learning, 1996, 24(2): 123-140

[16] 苏金树，张博锋，徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报，2006，17(9)：1848-1859

[17] Chawla N V, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 1-6

[18] 中国爬盟[EB/OL]. http://www.cnpameng.com, 2012