# 基于最大熵模型的导师 - 学生关系推测*

李勇军† 刘尊 于会

(西北工业大学计算机学院, 西安 710072)

(2013年4月16日收到; 2013年5月6日收到修改稿)

导师 - 学生关系是科研合作网络中重要的关系类型之一, 准确识别此类关系对促进科研交流与合作、评审回避等有重要意义. 以论文合作网络为基础, 依据学生发表论文时通常与导师共同署名的现象, 抽象出能够反映导师 - 学生合作关系的特征, 提出了基于最大熵模型的导师 - 学生关系识别算法. 利用 DBLP 中 1990—2011 年的论文数据进行实例验证, 结果显示: 1) 关系类型识别结果的准确率超过 95%; 2) 导师 - 学生关系终止时间的平均误差为 1.39 年. 该方法在识别关系时避免了特征之间相互独立的约束, 准确率优于其他同类识别算法, 且建模方法对识别社交网络中的其他关系类型也具有借鉴意义.

**关键词:** 社交网络, 关系识别, 最大熵模型, 特征选择

**PACS:** 89.75.Fb, 89.20.Ff, 05.90.+m **DOI:** 10.7498/aps.62.168902

## 1 引 言

科研合作网络是科研人员在科研合作过程中逐渐形成的合作关系网络. 通过对科研合作网络的研究可以更深入地认识科研人员的合作规律, 推动科研工作的合作、交流与创新. 在科研合作网络中, 不同类型的关系对科研合作的影响较大. 识别科研合作网络中的关系类型有助于准确地建立科研合作网的网络模型, 促进科研合作与协同创新. 导师 - 学生关系是科研合作中常见的一种关系类型. 与其他关系类型相比, 识别导师 - 学生关系在理解师承关系、学术团体挖掘、评审回避等方面有更大的实际意义.

近年来, 关系类型识别引起了学术界的关注. 最初的工作主要集中在识别朋友关系是否存在, 如 Bai 等 [1] 利用半局部相似索引 (semi-local similarity index) 研究了复杂网络中的链路预测; Backstrom 和 Leskovec[2] 提出了一种有监督的随机游走算法推测社交关系的强度; Leskovec 等 [3] 利用对数回归模型识别社交网络中的朋友关系或非朋友关系. 上述工作仅仅识别朋友关系是否存在, 并未区分关系类型. 后来的研究工作扩展到识别特定领域社交网络中的关系类型, 如 Diehl 等 [4] 通过学习到的排序函数识别公司中的经理 - 职员关系; Eagle 等 [5] 基于移动电话的通话关系挖掘了隐含在其中的通信模式, 并以此为基础识别隐含在网络中的关系类型. 这两类工作并不能直接应用到导师 - 学生关系识别中. Wang 等 [6] 挖掘了隐含在科研合作网络中的导师 - 学生关系, 该算法基于因子图建模, 推测结果中反例误判率比较理想, 但正例的准确率和覆盖率有待进一步提高. 近年来学者们也对社交网络中的复杂关系识别进行了研究, 如 Tang 等 [7] 跨越不同社交网络 (如邮件网络、科研合作网络等) 提出一种基于因子图的统一框架识别朋友关系类型, 其建模方法与文献 [6] 类似. Tang 等 [8] 利用人人网中已经标注出来的朋友类型和社交网络中的社团属性来识别朋友之间的关系. 但该方法假设同一个社区内人的关系类型相同, 并不完全符合实际情况. 另外该方法对已经明确标注出来的关系类型有较高的要求, 而现实中的科研合作网络中很少明确标注出关系类型. 可见, 除文献 [6] 外, 已有工作没有考虑科研合作网络的实际情况, 而不能直接应用到导师 - 学生关系识别中, 但对推测师生关系具有借

* 西北工业大学基础研究基金 (批准号: NPU-FFR-JC201257, JCY20130137) 资助的课题.

† 通讯作者. E-mail: lyj@nwpu.edu.cn

 

鉴意义.

为提高导师 - 学生关系识别的准确率和覆盖率, 本文以科研人员的论文合作网络为数据基础, 研究导师 - 学生关系识别方法. 在论文合作网络中, 学生发表的论文中通常会署导师的姓名, 而导师发表的论文中未必署学生的姓名; 导师发表的第一篇论文的时间早于学生; 导师发表论文的数量多于学生的数量等. 根据这些科研合作网络的特点, 抽象出具体的特征, 提出基于最大熵模型和特征选择 (maximum entropy model with feature selection, MEM-FS) 导师 - 学生关系识别算法, 实例验证的结果显示 MEM-FS 的准确性超过 95%. 算法 MEM-FS 的建模方法可以扩展到社交网络中的关系识别, 对研究社交网络上的信息传播 [9,10] 和复杂网络中的关键节点 [11] 识别具有现实意义.

## 2 问题描述

用二分网络 $\boldsymbol{G}=(E,A,P)$ 表示论文合作网络 (在下文中, 论文合作网络也称科研合作网络), 其中 $\boldsymbol{A}=\{a_1, a_2, \cdots, a_n\}$ 是网络中所有作者的集合, $\boldsymbol{P}=\{p_1, p_2, \cdots, p_m\}$ 是所有文章的集合, $\boldsymbol{E}=\{e_{ik}|1\leqslant i\leqslant n, 1\leqslant k\leqslant m, a_i$ 是 $p_k$ 的作者 $\}$ 是所有边的集合. 在网络 $\boldsymbol{G}$ 中, 作者之间不存在边, 但作者之间隐含了不同类型的关系, 比如导师 - 学生关系等. 本文工作主要是识别这些隐含在科研合作网络中的导师 - 学生关系.

**定义 1** 变量集合 $\boldsymbol{R}=\{r_{ij}|1\leqslant i\leqslant n, 1\leqslant j\leqslant n\}$ 表示作者之间的相互关系, 其中 $r_{ij}$ 表示作者 $a_i$ 和 $a_j$ 之间的关系.

**定义 2** 二元变量集合 $\boldsymbol{TMPY}=\{y_{ij}|1\leqslant i\leqslant n, 1\leqslant j\leqslant n\}$ 表示作者之间是否是导师 - 学生关系, 其中 $y_{ij}$ 对应 $r_{ij}$, 其关系如下:

$$y_{ij}=\begin{cases}1 & r_{ij}\text{是导师 - 学生关系}\\ 0 & \text{否则}\end{cases}. \tag{1}$$

文中的导师 - 学生关系特指学生读书期间的关系, 开始时间为学生入学时间, 终止时间为学生毕业时间. 为准确地描述特定时间段内的导师 - 学生关系, 扩展定义 2 中的二元变量为一个三元组.

**定义 3** 集合 $\boldsymbol{Y}=\{\{y_{ij}, st_{ij}, et_{ij}\}|1\leqslant i\leqslant n, 1\leqslant j\leqslant n\}$ 表示作者之间是否是导师 - 学生关系以及该关系的开始时间和终止时间, 其中 $st_{ij}$ 表示关系的开始时间, $et_{ij}$ 表示关系的终止时间.

基于上述的基本定义, 导师 - 学生关系推测问题定义如下.

**定义 4** 导师 - 学生关系推测问题 给定科研合作网络 $\boldsymbol{G}$, 推测集合 $\boldsymbol{R}$ 中的每对作者关系 $r_{ij}$ 是否是导师 - 学生关系以及关系存续时间段, 即 $\{y_{ij}, st_{ij}, et_{ij}\}$. 该问题可以定义为一个最大后验概率问题, 即

$$\hat{y}=\arg\max_{y}\boldsymbol{P}(y|\boldsymbol{G}), \tag{2}$$

其中 $\hat{y}=\{\hat{st}_{ij}, \hat{et}_{ij}, \hat{y}_{ij}\}$.

判断关系 $r_{ij}$ 是否是导师 - 学生关系实际上是个分类问题. 在网络 $\boldsymbol{G}$ 中可以找到很多用于分类的特征, 如每位作者的文章数量、合作的文章数量等, 但这些特征之间并非完全独立, 如两位作者合作的文章数量与每位作者的论文数量是关联的, 因此本文采用不要求特征独立的最大熵模型解决该问题.

## 3 基于最大熵模型的导师 - 学生关系推测

从定义 4 中可知, 解决导师 - 学生关系推测问题需要推测三个数值, 即是否是导师 - 学生关系 $y_{ij}$, 开始时间 $st_{ij}$ 以及终止时间 $et_{ij}$.

### 3.1 推测开始时间 $st_{ij}$

从科研合作网络提供的信息中, 很难准确地推测出导师 - 学生关系的真正开始时间. 本文用导师和学生共同署名发表的第一篇文章的时间作为关系的开始时间 $st_{ij}$. 一般来说, 学生发表第一篇论文的时间要晚于学生的入学时间, 因此推测的 $st_{ij}$ 比实际的关系开始时间滞后.

### 3.2 推测终止时间 $et_{ij}$

借鉴文献 [6] 的方法, 利用 Kulczinski 系数 (下文简称 Kulc 系数) 推测关系终止时间, 其定义如 (3) 式所示.

$$\mathrm{Kulc}_{ij}^{t}=\frac{\sum_{k\leqslant t}|pn_{ij}^{k}|}{2}\times\left(\frac{1}{\sum_{k\leqslant t}|pn_{i}^{k}|}+\frac{1}{\sum_{k\leqslant t}|pn_{j}^{k}|}\right), \tag{3}$$

其中, $|pn_i^k|$ 表示作者 $a_i$ 在第 $k$ 年发表的文章数量, $|pn_{ij}^k|$ 表示作者 $a_i$ 和 $a_j$ 在第 $k$ 年共同发表的文章数量. 文献 [12] 指出导师和学生发表文章的数量具

有较高的相关性, 随着导师和学生合作论文数量的增加, Kulc 系数也增加. 在导师 - 学生关系存续期间, 学生发表的论文通常署导师的名字, Kulc 系数会逐年增加; 学生毕业后发表的论文不再署导师名字, 此时 Kulc 系数会停止增加. 因此, 通过计算每对作者每年的 Kulc 系数, 根据该系数的变化可以判断关系的终止时间. 由此推测关系终止时间和开始时间一样也有滞后性, 因为论文从完成到发表有一定的时间间隔.

## 3.3 基于最大熵模型推测导师 - 学生关系

推测导师 - 学生关系的核心思想是, 学生通常和导师合作发表论文, 因此在学生发表的论文中, 其导师通常是作者之一; 而在导师发表的论文中, 学生未必是合作作者. 推测过程中不涉及第三位作者. 因此, (2) 式中导师 - 学生关系推测问题可以表示为

$$P(y_{ij}|G) = P(y_{ij}|G'_{ij}), \tag{4}$$

其中 $\boldsymbol{G}'_{ij}$ 是科研合作网络中作者 $a_i$ 和 $a_j$ 及他们的文章所构成的二分子图.

导师 - 学生关系推测结果受诸多因素影响, 如每位作者发表的文章数量、两位作者共同发表文章数量的比值等, 把所有这些影响因素记为集合 $\boldsymbol{X}$, 其中 $\boldsymbol{x} \in \boldsymbol{X}$ 表示某种影响因素. 为便于表示各种因素对推测结果的影响, 引入指示函数如下:

$$f(x,y) = \begin{cases} 1 & \text{如果}(x,y)\text{满足特定条件} \\ 0 & \text{否则} \end{cases}, \tag{5}$$

指示函数 $f$ 又称特征函数, 简称为特征. 在最大熵模型中, 多种因素对应多个特征. 利用最大熵模型推测导师 - 学生关系时, (4) 式所示的条件概率可改写为

$$p(y|G) = Z_\lambda(x)\exp\left(\sum\nolimits_i \lambda_i f_i(x,y)\right), \tag{6}$$

$$Z_\lambda(x) = 1\Big/\sum\nolimits_y \exp\left(\sum\nolimits_i \lambda_i f_i(x,y)\right), \tag{7}$$

其中, $Z_\lambda$ 是归一化因子, $\lambda_i$ 是权重因子, 表示特征 $f_i$ 的重要性.

影响推测导师 - 学生关系的因素可分为两大类: 作者特征 (author features) 和关系特征 (relationship features).

作者特征 $f_U(x,y)$ 在某种程度上反映一位作者具有导师或学生身份的可能性, 仅与单个作者相关. 例如当一位作者发表的论文数量少于 5 篇时, 该作者具有导师身份的概率会很低.

关系特征 $f_R(x,y)$ 指两位作者发表论文数量的相互关系, 是推测导师 - 学生关系的关键特征. 比如, 两位作者发表论文的比值、共同发表论文数量与各自发表论文数的比值等.

基于上述两个特征的定义, (6) 式可以进一步表示为

$$\begin{aligned} p(y|x) &= p(y|G) \\ &= Z_\lambda(x)\exp\left\{\sum_i^{K(A)} \lambda_i f_{Ai}(x,y) \right. \\ &\quad \left. + \sum_i^{K(R)} \lambda_i f_{Ri}(x,y)\right\}, \end{aligned} \tag{8}$$

其中 $K(\boldsymbol{A})$ 是作者特征的数量, $K(\boldsymbol{R})$ 是关系特征的数量.

假设学习集合为 $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_n, y_n)\}$, $p'(x,y)$ 表示样本 $(x,y)$ 的经典概率. 特征函数 $f$ 和因素 $x$ 的经典概率分布分别为

$$p'(f) \equiv \sum\nolimits_{x,y}(p'(x,y)\cdot f(x,y)), \tag{9}$$

$$p'(x) \equiv \sum\nolimits_y p'(x,y). \tag{10}$$

参数集合 $\boldsymbol{\lambda} = \{\lambda_i, i = 1, 2, \cdots, K(\boldsymbol{A}) + K(\boldsymbol{R})\}$ 可通过学习已有的样本集合求解. 在学习过程中, 为防止参数被过度训练, 通常为参数假设一个先验分布, 本文采用的分布是高斯分布.

$$p(\lambda_i) = \frac{1}{\sqrt{2\pi}\delta}\exp\left\{-\frac{(\lambda_i-\mu)^2}{2\delta^2}\right\}. \tag{11}$$

待学习的参数总量 $(K(\boldsymbol{A}) + K(\boldsymbol{R}))$ 记为 $k$, 学习参数的正则对数似然函数为

$$\begin{aligned} L(\lambda) &= \sum_{x,y}\left\{p'(x,y)\cdot\log p(y|x)\right\} - \sum_i \log p(\lambda_i) \\ &= \sum_{x,y}\left\{p'(x,y)\left(\sum_{i=1}^{K(A)} \lambda_i f_{Ai}(x,y) \right.\right. \\ &\quad \left.\left. + \sum_{i=1}^{K(R)} \lambda_i f_{Ri}(x,y) - \log Z_\lambda(x)\right)\right\} \\ &\quad - \sum\nolimits_{i=1}^{k} \frac{(\lambda_i-\mu)^2}{2\delta^2} - k\log\sqrt{2\pi}\delta. \end{aligned} \tag{12}$$

似然函数 $L(\boldsymbol{\lambda})$ 是凸函数, 利用

$$\frac{\partial L(\lambda)}{\partial\lambda} = 0, \tag{13}$$

求解最优参数值, 但实际中很难找到一个解析解, 一般采用基于梯度的数值优化算法进行求解,

目前常用的算法是算法是约束 Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) 算法 [13].

### 3.4 算法描述

基于最大熵模型的导师 - 学生推测算法主要步骤描述如下:

**输入** 训练 (学习) 数据集、作者特征 $f_U(x,y)$ 和关系特征 $f_R(x,y)$

**输出** 特征函数的权重参数 $\boldsymbol{\lambda}=\{\lambda_i,\ i=1,2,\cdots,k\}$

**步骤 1** 从训练集中任取两位作者, $a_i$ 和 $a_j$, 推测开始时间 $st_{ij}$;

**步骤 2** 利用 (3) 式计算 $a_i$ 和 $a_j$ 每年的 Kulc 系数, 根据 Kulc 系数推测 $et_{ij}$;

**步骤 3** 如果 $et_{ij}-st_{ij}>1$, 作者对 $(a_i,a_j)$ 放入潜在关系数据集合 $\boldsymbol{Q}$ 中;

**步骤 4** 重复步骤 1—3 直到训练集中所有作者对被遍历;

**步骤 5** 从 $\boldsymbol{Q}$ 任取出一对作者 $(a_i,a_j)$, 按照特征函数 $f_U(x,y)$ 和 $f_R(x,y)$ 的定义从作者 $a_i$ 和 $a_j$ 发表论文列表中获取相应的特征值, 并计算 $p'(x,y)$;

**步骤 6** 重复步骤 5 直到 $\boldsymbol{Q}$ 为空;

**步骤 7** 获取的特征值 $\{(x_1,\ y_1),(x_2,\ y_2),\ \cdots,(x_n,\ y_n)\}$及其相应的经典概率 $p'(x,y)$ 代入 (12) 和 (13) 式中, 利用 L-BFGS 算法求解 $\boldsymbol{\lambda}=\{\lambda_i,\ i=1,2,\cdots,k\}$.

得到模型参数 $\boldsymbol{\lambda}=\{\lambda_i,\ i=1,2,\cdots,k\}$后, 基于最大熵模型的导师 - 学生关系推测模型训练完成, 可用于推测导师 - 学生关系. 由 (8) 式可知, 推测结果是一个概率值. 如果作者 $a_i$ 和 $a_j$ 是导师 - 学生关系的概率大于给定阈值, 则推测算法输出 $a_i$ 是 $a_j$ 的导师, 起止时间分别为 $st_{ij}$ 和 $et_{ij}$. 否则, 作者 $a_i$ 和 $a_j$ 是合作关系而不是导师 - 学生关系.

## 4 实例验证

以 1990—2011 年的 DBLP 数据为测试集, 构建科研合作网络, 检验 MEM-FS 的性能. 为减小分类算法的收索空间, 提高算法的性能, 需对测试集进行预处理. 首先去掉发表文章数量少于 3 篇的作者, 因为即使采用人工的方法也很难准确地识别出这些作者的导师; 其次, 如果两位作者在发表论文的时间上不存在交集, 显然不可能是师生关系, 这两位作者的关系也从候选集合中过滤掉. 通过统计发现经过数据预处理, 师生关系的候选集合大小为原来的约 10%. 过滤后的数据集包含 996427 位作者和 1656588 篇论文. 为检验算法的准确性, 以文献 [6] 中的 MAN, MathGP 和 AIGP 三个数据集的并集作为验证数据集, 其中 MAN 数据集通过查询导师的个人主页人工构建出, MathGP 和 AIGP 则分别是通过爬取 Mathematics Genealogy 和 AI Genealogy 的相关页面获得. MAN 数据集进一步分为三个子数据集: Teacher, PhD 和 Colleague. Teacher 集中包含了在读和已经毕业的学生、导师以及学生的毕业时间; PhD 中仅包含在读学生和导师的信息; Colleague 是个反例数据集, 是合作者对的集合, 不包含导师 - 学生关系.

为直观地比较推测结果, 本文采用分类算法常用的评价指标: 受试者特征曲线 (ROC) 曲线中的真阳性率 (true positive rate, TPR) 和假阳性率 (false positive rate, FPR) 以及精确度 (accuracy, ACC) 三个指标.

MEM-FS 算法采用的候选特征集见表 1.

### 4.1 不同特征组合对结果的影响

导师 - 学生关系推测结果受多种因素影响, 且不同因素的影响大小也不尽相同. 以表 1 中的候选因素为例, 对 7 种不同特征集进行排列, 用实例验证不同因素对推测结果的影响, 7 种不同特征排列如表 2 所示. 在训练集中, 作者 $a_i$ 是学生, $a_j$ 是导师. 按照此特征排列准备测试数据集, 推测作者 $a_j$ 是否是作者 $a_i$ 的导师.

表 1 特征列表

| 特征 | 特征描述 |
|---|---|
| $\lvert pn_i\rvert, \lvert pn_j\rvert$ | 作者发表文章的数量 |
| $\lvert pn_i\rvert/\lvert pn_j\rvert, \lvert pn_j\rvert/\lvert pn_i\rvert$ | 两位作者发表文章数量的比值 |
| $\lvert pn_i\cap pn_j\rvert/\lvert pn_j\rvert, \lvert pn_i\cap pn_j\rvert/\lvert pn_i\rvert$ | 两位作者合作论文数量与各自论文数量的比值 |
| $st_i-st_j, st_j-st_i$ | 两位作者第一篇文章见刊时间的差值 |

表 2 不同特征排列列表

| 编号 | 特征排列 |
|---|---|
| 1 | $\|pn_i\|$, $\|pn_j\|$, $\|pn_i\|/\|pn_j\|$, $\|pn_i \cap pn_j\|/\|pn_i\|$, $\|pn_i \cap pn_j\|/\|pn_j\|$, $st_i - st_j$ |
| 2 | $\|pn_i\|$, $\|pn_j\|$, $\|pn_i\|/\|pn_j\|$, $\|pn_i \cap pn_j\|/\|pn_i\|$, $\|pn_i \cap pn_j\|/\|pn_j\|$ |
| 3 | $\|pn_i\|/\|pn_j\|$, $\|pn_i \cap pn_j\|/\|pn_i\|$, $\|pn_i \cap pn_j\|/\|pn_j\|$, $st_i - st_j$ |
| 4 | $\|pn_i\|/\|pn_j\|$, $\|pn_i \cap pn_j\|/\|pn_i\|$, $\|pn_i \cap pn_j\|/\|pn_j\|$ |
| 5 | $\|pn_i\|/\|pn_j\|$, $\|pn_i \cap pn_j\|/\|pn_i\|$ |
| 6 | $\|pn_i\|/\|pn_j\|$, $\|pn_i \cap pn_j\|/\|pn_j\|$ |
| 7 | $\|pn_i \cap pn_j\|/\|pn_i\|$ |

基于不同特征排列的推测结果见表 3. 从 ACC 上讲, 推测过程中涉及的因素越多, 结果越准确. 特征排列 1 的 ACC 最高, 而特征排列 6 和 7 的 ACC 均未超过 90%. 此外, 特征排列 2 中包含 5 种不同特征, 仅比特征排列 1 少了特征 $(st_i - st_j)$, 但其对应的推测结果比特征排列 1 低较多, 可见特征 $(st_i - st_j)$ 在推测过程具有较重要的作用. 从 FPR 上也可以看出, 特征排列 2 的 FPR 比特征排列 1 高了 10%多, 说明推测结果中反例的误判率比特征排列 1 对应的结果高很多.

表 3 不同特征排列对应的推测结果

| 特征排列编号 | TPR/% | FPR/% | ACC/% |
|---|---|---|---|
| 1 | 97.99 | 19.08 | 95.51 |
| 2 | 96.31 | 29.61 | 92.54 |
| 3 | 97.76 | 19.74 | 95.22 |
| 4 | 98.55 | 22.37 | 95.51 |
| 5 | 99.44 | 32.89 | 94.74 |
| 6 | 90.49 | 81.58 | 80.02 |
| 7 | 1 | 1 | 85.47 |

特征排列 5 和 6 都仅包含两个特征, 差异在于特征排列 5 中包含特征 $(|pn_i \cap pn_j|/|pn_i|)$, 其物理意义是两位作者共同发表论文的数量与学生发表论文数量的比值, 而特征排列 6 包含的特征是两位作者共同发表论文的数量与导师发表论文数量的比值. 而这两种特征排列对应的推测结果差异较大, 从 TPR, FPR 和 ACC 三个指标上看, 特征排列 5 均优于特征排列 6, 说明特征 $(|pn_i \cap pn_j|/|pn_i|)$ 在推测过程中比 $(|pn_i \cap pn_j|/|pn_j|)$ 的作用大. 而表 3 也显示, 单一特征 $(|pn_i \cap pn_j|/|pn_i|)$ 的误判率是 100%, 说明即使很重要的特征, 单独用于推测过程中也是不可行的.

FPR 随着特征数量的减少会增加, 即推测过程中用到的因素越少结果中误判的概率就越高; 而对于 TPR, 表面上 TPR 随着特征数量的减少而略微增加, 即推测结果中查全率略微增加, 但 TPR 的增加是以 FPR 的增加为代价的. 可见单独依据 FPR 或 TPR 很难评价算法的性能. 图 1 给出了 7 种不同特征排列对应的 TPR 和 FPR 在 ROC 空间中的位置. 从图 1 中可以看出, 特征排列 1 的推测效果最优, 特征排列 7 的效果最差. 图 1 显示的结果和依据 ACC 排列的结果类似.

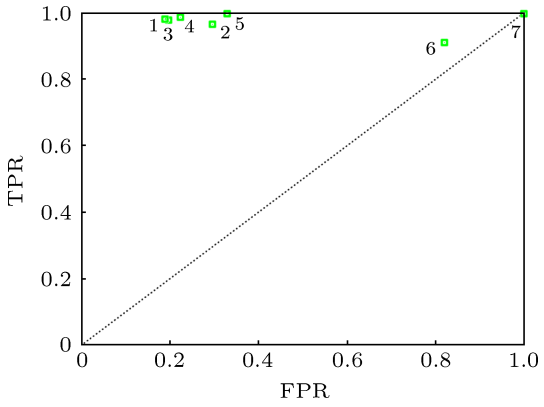


图 1 不同特征排列的推测结果比较

算法推测结果的准确性还体现在 $st_{ij}$ 和 $et_{ij}$ 的推测结果上, 如前所述 $st_{ij}$ 的实际时间很难获得, 本文仅评价 $et_{ij}$ 的推测效果. $et_{ij}$ 推测结果的平均误差是 1.39, 中值是 1. 推测误差可以解释为论文完成到出版的时间.

## 4.2 训练集大小对结果的影响

直观上训练集越大其效果会越好, 但随着训练集的增加, 训练时间也会变长. 理想的状况是, 选择一个大小合适的训练集, 并能取得最佳的训练效果. 表 4 给出了在不同大小数据集上, 算法 MEM-FS 的推测效果. 在推测中, 特征采用表 2 中的特征排列

1, 数据全集包含 900 条正例和 150 条反例. 在缩减数据集时, 正例和反例均按比例缩减.

表 4 不同大小的训练集对应的推测结果

| 编号 | 数据集大小 | TPR/% | FPR/% | ACC/% |
|---|---|---|---|---|
| 1 | 全集 | 97.99 | 19.08 | 95.51 |
| 2 | 全集的 8/10 | 97.54 | 26.97 | 93.98 |
| 3 | 全集的 6/10 | 98.88 | 29.61 | 94.74 |
| 4 | 全集的 4/10 | 98.77 | 34.87 | 93.88 |
| 5 | 全集的 2/10 | 98.10 | 31.58 | 93.79 |
| 6 | 全集的 1/10 | 96.09 | 44.08 | 90.25 |

从表 4 中可以看出, 随着数据集缩减, ACC 降低, 而 TPR 略微增加. FPR 相对于 TPR 和 ACC 变化最为明显, 主要原因是反例集本身就小, 缩减以后严重影响了学习效果, 使得 ACC 降低. 保持反例集的大小不变, 变化正例集的大小, 研究数据集大小对推测结果的影响, 实验结果如表 5 所示.

表 5 不同大小的正例集对应的推测结果

| 编号 | 数据集大小 | TPR/% | FPR/% | ACC/% |
|---|---|---|---|---|
| 1 | 正例全集 | 97.99 | 19.08 | 95.51 |
| 2 | 正例的 8/10 | 97.54 | 19.74 | 95.03 |
| 3 | 正例的 6/10 | 96.87 | 20.39 | 94.36 |
| 4 | 正例的 4/10 | 95.41 | 19.08 | 93.31 |
| 5 | 正例的 2/10 | 93.40 | 15.79 | 92.07 |
| 6 | 正例的 1/10 | 84.68 | 9.87 | 85.47 |

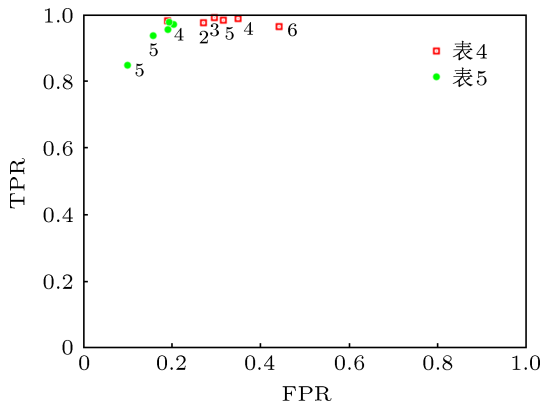


图 2 正反例比例对推测结果的影响

从表 5 中可以看出, 在保持反例大小不变的情况下, 随着正例规模缩小, TPR 和 ACC 都降低, 而 FPR 也呈现降低趋势. 另外, 比较表 4 和表 5 发现, 在正例比例相同的情况下, 表 4 中 ACC 值比表 5 中的要高. 由此可以说明训练效果与训练集中正例和反例的比例有关系. 图 2 给出了表 4 和表 5 中的 TPR 和 FPR 在 ROC 空间中的位置, 从图 2 中可以直观地比较两种情况下的推测结果. 表 4 中的 TPR 一般要高于表 5 中的对应值, 而表 5 中 FPR 要低于表 4 中的相应值. 在表 4 对应的训练集中, 正例和反例的比例是不变的, 而表 5 对应的比例是逐渐减小, 所以 FPR 逐渐降低.

## 4.3 与同类算法的对比结果

表 6 列出了四种不同算法的推测结果, 其中 TPFG[6] 是基于因子图的推测算法, C-SVM 算法是基于支持向量机的算法, 其核函数采用 RBF, Baseline 算法是自定义的基准算法, 是其余三种算法的一部分, 在推测出两两作者的关系起止时间后, 从所有合作者中选择合作次数最多的作者作为导师.

表 6 与 Baseline, TPFG 的推测结果对比

| 算法 | TPR/% | FPR/% | ACC/% |
|---|---|---|---|
| MEM-FS | 97.99 | 19.08 | 95.51 |
| C-SVM | 88.37 | 25.00 | 86.42 |
| Baseline | 74.19 | 3.93 | 77.37 |
| TPFG | 67.09 | 1.08 | 71.69 |

表 6 显示, MEM-FS 在三项指标上均优于 C-SVM 算法, 而与 Baseline 和 TPFG 算法相比, 在 TPR 和 ACC 上具有优势, 但 FPR 比 Baseline 和 TPFG 算法要明显高. MEM-FS 的 ACC 是以较高的 FPR 为代价, 也就是用较高的误判率换取准确性. 表 6 很难说明 MEM-FS 比其他两种算法的推测效果好. 为了更好地说明 MEM-FS 的性能, 调整 MEM-FS 的判断阈值, 使得 FPR 也降到 1%左右, 再与 TPFG 算发做比较, 调整后的结果见表 7.

表 7 与 TPFG 在近似 FPR 情况下的推测结果对比

| 算法 | TPR/% | FPR/% | ACC/% |
|---|---|---|---|
| MEM-FS | 71.47 | 1.31 | 75.43 |
| TPFG | 67.09 | 1.08 | 71.69 |

从表 7 中可以看出, 在近似 FPR 的情况下, MEM-FS 的推测效果略好于 TPFG. 二者之间的差异除算法本身原因外, 也可能由以下原因引起: 1) MEM-FS 和 TPFG 算法采用的训练集和测试集略有不同, 可能会影响到算法效果; 2) MEM-FS 采用的测试集仅有 152 个反例, 当 FPR 为 1.31%时, 误判的反例个数为 2, 具有偶然性.

MEM-FS 和 TPFG 算法在推测导师 - 学生关系时, 性能表现在不同方面: 利用 MEM-FS 可以在推断正例时达到较高的准确率和覆盖率, 而在反例方面却具有较高的误判率; 而 TPFG 在反例方面具有较低的误判率, 但在正例方面具有较低的正确率. TPFG 在推测导师 - 学生关系方面比 MEM-FS 相对谨慎. 从 ACC 上讲, MEM-FS 明显优于 TPFG 算法.

## 5 结 论

本文研究了论文合作网络中导师 - 学生关系的识别问题, 并将该问题细分为推测关系的开始时间、终止时间以及关系类型识别三个子问题. 在论文合作网络中, 学生发表的论文通常会署导师的姓名, 反之未然. 依据上述事实, 关系开始时间推测为导师和学生发表第一篇文章的时间; 基于导师和学生每年发表论文数量计算每年的 Kulc 系数, 根据逐年 Kulc 系数的变化推测关系的终止时间; 通过分析科研合作网络的特点, 提取出能够反映导师 - 学生关系的特征, 建立基于最大熵模型和特征选择的关系识别模型, 推测隐含在论文合作网络中的导师 - 学生关系. 利用 DBLP 中的数据集进行实例验证, 实验表明: 1) 关系终止时间的平均误差为 1.39 年, 误差产生的原因主要是论文出版时间滞后完成时间; 2) 关系类型识别结果的准确率超过 95%, 明显优于同类的识别算法.

感谢美国伊利诺伊大学香槟分校 (University of Illinois at Urbana-Champaign, UIUC) 的 Chi ZHANG 共享 TPFG 算法的结果和验证数据集.

[1] Bai M, Hu K, Tang Y 2011 *Chin. Phys.* B **20** 12

[2] Backstrom L, Leskovec J 2011 *Proceedings of the 4th ACM International Conference on Web Search and Data Mining* Hong Kong, China, February 9–12, 2011 pp635–644

[3] Leskovec J, Huttenlocher D P, Kleinberg J M 2010 *Proceedings of 19th International World Wide Web Conference* Raleigh, USA, April 26–30, 2010 pp641–650

[4] Diehl C P, Namata G, Getoor L 2007 *Proceedings of Twenty-Second Conference on Artificial Intelligence* Vancouver, Canada, July 22–26, 2007 pp546–552

[5] Eagle N, Pentland A S, Lazer D 2009 *Proc. Nat. Acad. Sci. U. S. A* **106** 36

[6] Wang C, Han J, Jia Y, Tang J, Zhang D, Yu Y, Guo J 2010 *Proceedings of 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Washington D.C., USA, July 24–28, 2010 pp203–212

[7] Tang J, Lou T, Kleinberg J 2012 *Proceedings of the 5th ACM International Conference on Web Search and Data Mining* Seattle, USA, February 8–12, 2012 pp743–752

[8] Tang S, Yuan J, Mao X, Li X, Chen W, Dai G 2011 *Proceedings of 30th IEEE International Conference on Computer Communications* Shanghai, China, April 10–15, 2011 pp1661–1669

[9] Zhang Y Ch, Liu Y, Zhang H F, Cheng H, Xiong F 2012 *Acta Phys. Sin.* **60** 050501(in Chinese) [张彦超, 刘云, 张海峰, 程辉, 熊菲 2012 物理学报 **60** 050501]

[10] Gu Y R, Xia L L 2012 *Acta Phys. Sin.* **61** 238701 (in Chinese) [顾亦然, 夏玲玲 2012 物理学报 **61** 238701]

[11] Yu H, Liu Z, Li Y J 2013 *Acta Phys. Sin.* **62** 020204 (in Chinese) [于会, 刘尊, 李勇军 2013 物理学报 **62** 020204]

[12] Wu T Y, Chen Y G, Han J W 2010 *Data Min. Knowl. Disc.* **21** 3

[13] Byrd R H, Nocedal J, Schnabel R B 1994 *Mathematical Programming* A, B **63** 4

# Advisor-advisee relationship identification based on maximum entropy model*

Li Yong-Jun† Liu Zun Yu Hui

( *School of Computer, Northwestern Polytechnical University, Xi'an* 710072, *China* )

( Received 16 April 2013; revised manuscript received 6 May 2013 )

**Abstract**

Research collaboration network has become an essential part in our academic activities. We can keep or develop collaboration relationships with other researchers or share research results with them within the research collaboration network. It is well generally accepted that different relationships have essentially different influences on the collaboration of researchers. Such a scenario also happens in our daily life. The advisor-advisee relationship plays an important role in the research collaboration network, so identification of advisor-advisee relationship can benefit the collaboration of researchers. In this paper, we aim to conduct a systematic investigation of the problem of indentifying the social relationship types from publication networks, and try to propose an easily computed and effective solution to this problem. Based on the common knowledge that graduate student always co-authors his papers with his advisor and not vice versa, our study starts with an analysis on publication network, and retrieves these features that can represent the advisor-advisee relationship. According to these features, an advisor-advisee relationship identification algorithm based on maximum entropy model with feature selection is proposed in this paper. We employ the DBLP dataset to test the proposed algorithm. The results show that 1) the mean of deviation of estimated end year to graduation year is 1.39; 2) the accuracy of advisor-advisee relationship identification results is more than 95%, and it is better than those of other algorithms obviously. Finally, the proposed algorithm can be extended to the relationship identification in online social network.

**Keywords:** social network, relationship identification, maximum entropy, feature selection

**PACS:** 89.75.Fb, 89.20.Ff, 05.90.+m **DOI:** 10.7498/aps.62.168902

* Project supported by the Fundamental Research Foundation of Northwestern Polytechnical University, China (Grant Nos. NPU-FFR-JC201257, JCY20130137).

† Corresponding author. E-mail: lyj@nwpu.edu.cn