基于监督联合去噪模型的社交网络链接预测

郝占刚^{1*}, 章伟雄², 陈政²

1. 山东工商学院工商管理学院, 烟台 264005, 中国

2. Department of Computer Science and Engineering, Washington University, St. Louis 63130, USA

* 通信作者. E-mail: zghao2000@sina.com

收稿日期: 2017-03-29; 接受日期: 2017-06-26; 网络出版日期: 2017-11-10

国家自然科学基金 (批准号: 71672103, 61572104, 71501113) 和山东省自然科学基金 (批准号: ZR2017MG022) 资助项目

摘要 社交网络链接预测能够获取丢失链接的重要信息, 在很多领域都有应用. 现有链接预测算法或因无法充分利用信息或因无法捕获全部特性而在链接预测精度上有所欠缺. 为提高链接预测精度, 本文提出一种新的算法——监督联合去噪模型 (supervised joint denoising model, SJDM), 该算法把链接预测问题看作一个有监督的矩阵“去噪”问题. 为了能够获取理想社交网络矩阵, 该算法综合利用现有用户特征信息和链接信息去训练一个映射函数. 该函数能够通过现有社交网络链接矩阵获取理想社交网络链接矩阵, 完成链接预测. 获取的理想社交网络链接矩阵是一个权重矩阵, 采用该矩阵进行社区发现能取得比采用原矩阵更佳的效果. 本文采用 5 个真实社交网络对 SJDM 算法进行了验证, 实验结果证明了该算法的有效性.

关键词 社交网络分析, 预测, 监督学习, 社区网络, 矩阵“去噪”, 弱大数定律

1 引言

在线社交网络 (online social network, OSN) 已深入大众生活, 成为人们重要的沟通及交友工具. 同时, 在线社交网络在政治经济中也发挥着越来越重要的作用, 比如 2016 年美国总统选举. 链接预测 (link prediction) 研究是社交网络研究的一项重要内容, 其应用很多^[1], 比如朋友推荐^[2]、寻找潜在合作者^[3,4]、寻找跨领域合作伙伴^[5]、在学术社交网络中寻找专家或者合作作者等^[6,7]. 本文所研究的链接预测主要指利用已知社交网络信息预测未知链接. 如图 1 所示, 图 N 为已知链接 (实线), 图 $\bar{N}$ 为未知链接 (虚线), 本文链接预测研究是指利用图 N 的已知信息去预测图 $\bar{N}$ 中的未知链接.

已有较多学者对社交网络链接预测问题进行了研究^[1]. 有些学者的研究基于节点对之间的相似性^[8~10], 这类研究基于一个基本的思想: 节点对之间越是相似, 它们之间存在链接的可能性越大. 有些学者基于社交网络的拓扑结构进行研究^[11~13], 即使最简单的没有节点和边属性的网络, 其网络结构、

引用格式: 郝占刚, 章伟雄, 陈政. 基于监督联合去噪模型的社交网络链接预测. 中国科学: 信息科学, 2017, 47: 1551–1565, doi: 10.1360/N112016-00262
Hao Z G, Zhang W X, Chen Z. Link prediction in online social networks based on supervised joint denoising model (in Chinese). Sci Sin Inform, 2017, 47: 1551–1565, doi: 10.1360/N112016-00262

图 1 链接预测问题示例图

Figure 1 Example of link prediction

网络特征也能够对节点相似性的研究提供帮助, 而且现实中很多社交网络是很复杂的. 有些学者基于社会理论对该问题进行了研究^[14, 15], 这些学者使用经典的社会理论, 比如社区理论、强弱联系理论、同质性理论等对链接预测问题进行研究. 有些学者基于特征、内在属性和外在信息提出了基于学习的链接预测方法^[16, 17].

Chen 等^[18] 基于矩阵“去噪”理论提出了 mDA 算法用于领域适应学习研究, 其也可用于社交网络链接预测. 但该算法仅利用网络结构信息, 而无法利用用户特征信息, 这也降低了其链接预测精度. 为提高链接预测精度, 本文提出一种新方法——监督联合去噪模型 (supervised joint denoising model, SJDM), 该方法把链接预测问题看作一个有监督矩阵“去噪”问题. 与 mDA 相比, 新模型是专门针对链接预测问题而构建的一个全新模型, 该模型可以充分利用网络结构信息和用户特征信息, 具有更高的预测精度. 该方法把已有社交网络看作一个不完整网络, 这种不完整是相对于理想社交网络而言 (理想社交网络指充满所有链接的网络, 所有链接指所有两个用户之间应该建立的链接, 建立链接的两个用户在社交网络中有可能成为朋友, 当然并不是所有用户之间都存在链接). 这个思路类似于深度学习中使用的降噪自动编码器. 在降噪自动编码器方法中, 去噪方法被广泛地使用在深度网络的初始化中. 其思路是希望深度网络中的“编码”层可以更好地概括拥有缺失信息的数据, 并由“解码”层来重构完整的数据. 而在社交网络中如果把最终需要预测的网络看作完整的数据, 那么当前观察到的网络就是拥有缺失信息的网络, 其中很多链接没有被观察到. 该方法主要是构造一个函数, 该函数能够把已有社交网络映射到理想社交网络. 在社交网络中, 一个用户和另一个用户是否建立链接受以下两个因素的影响很大: 一是两个用户之间的相似性, 两个用户的相似性越高, 二者之间建立链接的可能性越大, 比如同属某个城市、同属某个大学等; 二是两个用户和其他用户的链接, 比如用户 u_1 和 u_2 , 如果两个用户之间拥有共同朋友 u_3 , 二者就有可能建立相互链接. 尽管这些数据可能存在“噪声”, 但这些数据对用户之间是否建立链接仍然起着至关重要的作用. 我们希望能够充分利用用户信息和网络拓扑结构信息建立一个映射函数, 通过该函数去获取理解社交网络.

为构造所需函数, 本文把这个问题看作一个权重矩阵范数最小化问题, 通过解这个最小化问题, 得到所需函数. 通过该函数, 能够获取一个带权重的理想社交网络, 基于该社交网络的社区发现效果要优于使用原社交网络. 本文方法主要贡献如下:

(1) 在准确性方面, 通过在 4 个较小规模和 1 个大规模真实社交网络数据集上的实验表明, 无论是在小规模社交网络还是在大规模社交网络中, SJDM 方法都能够显著地提高社交网络链接预测的精度.

(2) 在计算效率方面, SJDM 方法本质上是一种基于局部的算法, 比起基于全局的算法, 它在空间复杂度和时间消耗上都要更小. SJDM 算法实际运算中可以采用低秩近似, 低秩矩阵能够节省大量内

存空间, 提高计算效率.

(3) 可以提供更准确的朋友推荐, SJDM 算法获取的加权矩阵能够提供任意一个用户 u_r 和其他所有用户的链接强度排名, 可以据此得出推荐排名. u_r 和排名靠前用户成为朋友的概率显然更高, 从而提高朋友推荐的准确性. 而加权矩阵中的负值则意味着两个用户之间不应该存在链接, 在做推荐时一定要避免.

(4) 在应用拓展方面, 通过 SJDM 方法获取的加权理想矩阵可以用于社交网络中的社区发现, 能取得更佳效果.

2 相关工作

现有社交网络链接预测算法可以分为两类: 基于特征的模型和隐变量模型. 下面对这两类算法在近几年的文献做简要综述.

基于特征的模型. Bhattacharyya 等基于用户关键词的相似性对用户之间的相似度进行评估; 他们定义了一个“森林模型”, 并通过该模型在语义相关性基础上来链接相关关键词; 基于该模型, 他们定义了两种不同类型的函数来量化用户之间的相似性^[1,8]. Papadimitriou 等^[19] 定义了一个新的节点相似度衡量方法, 该方法综合利用网络的局部和全局特征. Symeonidis 等^[20] 采用谱聚类的方法对社交网络的链接预测进行了研究; 他们通过降维的方法获得一个更小的矩阵, 采用该矩阵能够提高预测效果; 他们不但计算了类内节点之间的相似度而且计算了不同类之间的节点相似度. Bastani 等^[21] 提出模糊模型对社交网络链接预测进行研究, 他们基于粒计算和模糊逻辑提出了两种新的相似指数, 并且基于多个链接预测模型提出了一个新的复合模型, 该模糊模型能够通过更好的网络特征表示来提高链接预测精度. Bliss 等^[22] 提出一个基于领域相似性和节点相似性的线性组合模型, 并采用自适应协方差矩阵进化策略对该模型权重进行优化, 该模型能够较好地进行链接预测. Bayrak 等^[23] 针对基于位置的社交网络提出一个基于语境特征的链接预测方法, 该方法能够提高预测的精度. Xie 等^[24] 提出了一种联合预测特征模型, 该模型考虑了用户的特征和网络结构特征, 并把链接预测问题转化为一个分类问题进行研究.

隐变量模型. Zhu 等^[25] 提出一个混合时间序列链接预测模型, 把社交网络用户的时间因素和整个社交网络的拓扑结构相结合进行社交网络链接预测. Liu 等^[26] 首先构建了一个特征转化方法, 该方法能够把原有特征大大简化, 简化后特征将比原特征更能提高链接预测精度; 其次构建基于受限 Boltzmann 机的深度学习算法进行链接预测, 这是个无监督算法, 其仅使用小样本训练即可取得良好效果. Gong 等^[27] 对社交-属性网络方法进行了扩展, 该算法能够基于网络结构和节点属性对社交网络链接进行预测, 他们把该算法与几种监督和无监督链接预测算法相结合来对社交网络链接进行预测. Panwar 等^[28] 提出一种启发式算法进行社交网络的链接预测, 其研究的社交网络主要为研究者或作家社交网络. He 等^[29] 提出一种基于 OWA (ordered weighted averaging) 操作算子的链接预测方法, 在该文中采用 3 种不同的 OWA 算子进行链接预测, 取得了良好的效果. Sherkat 等^[30] 提出一种基于蚁群算法的社交网络链接预测方法, 蚁群算法主要用来发现特殊的子图, 通过对这些子图进化的研究去预测社交网络的新链接. Nguyen-Thi 等^[31] 基于 boosting 学习算法提出一种新的迁移学习框架, 该框架包括一个包含 RBF 核函数的 SVM 分类器; 在该框架中, 他们使用两类特征对社交网络链接预测进行研究.

对基于特征的模型而言, 特征是其进行链接预测的根基, 但现实中有些特征不易获取, 获取的特征也可能存在“噪声”. SJDM 算法尽管也需要用到特征, 但需要的特征数可以比较少, “噪声”特征对

图 2 算法流程示意图
Figure 2 Flow chart of SJDM algorithm

其影响很小, 而且特征只是算法所需信息的一部分. 社交网络往往呈现“同质性 (朋友的朋友)”和“统计上相同”两种特性^[32], 隐变量模型很难同时获取这两种特性, 而 SJDM 算法则可以同时获取, 从而提高预测的精度.

3 SJDM 算法

本文使用大写字母 X, Y 表示矩阵, 使用小写字母 x, y 表示列向量, $(X)_{ij}$ 或 X_{ij} 表示矩阵 X 的第 i 行第 j 列元素. X 在本文中表一个现有的 $n \times n$ 社交网络链接矩阵, 在该矩阵中只有一些现存链接, 非常稀少, 是一个 0, 1 稀疏矩阵. 如果 $X_{ij} = 1$, 表示第 i 个社交网络用户和第 j 个社交网络用户之间存在链接, 否则表示两个用户之间不存在链接, 在该矩阵中存在大量的数字 0, 即大量用户之间不存在链接. 数字 0 在社交网络中表示两种含义: 一是两个用户之间应该存在链接, 但现有的社交网络并没有反映出来; 二是两个用户之间不应该存在链接, 但现有社交网络也无法反映. 因此, 本文希望能够建立一个新的链接矩阵 Y , 该矩阵包含所有应该建立的链接, 也可以清晰地表示不应该建立的链接. 为了能够顺利获取该矩阵 Y , 本文把社交网络链接预测问题看作一个有监督机器学习问题. 通过解该问题, 获取一个函数 $h \in \mathcal{H}$, 该函数能够把已有矩阵 X 映射到预测矩阵 Y , 即 $Y = h(X)$. 算法流程示意图如图 2 所示.

3.1 问题构建

因为矩阵 X 是不完整矩阵, 缺少很多链接, 本文希望基于该矩阵去发现缺失的链接. 相对于 Y 而言, X 中包含的链接很少, 其可以被看作是从 Y 中随机抽取若干链接后所形成的链接矩阵. 我们希望训练一个映射函数 h , 通过该函数能够把“被破坏的”矩阵 X 映射到目标矩阵 Y . 因为目标矩阵 Y 是希望获取的, 是未知的, 因此直接通过矩阵 X 获取映射函数 h 是非常困难的. 为获取该函数, 本文使用一个数据驱动的有监督判别学习算法来训练该函数. 首先对现有矩阵 X 进行进一步的“破坏”, 可

以得到一个更深层次“被破坏的”矩阵 $\tilde{X}$. 然后希望通过 $\tilde{X}$ 来训练函数 h , 并使得 h 能够把矩阵 $\tilde{X}$ 映射到矩阵 X . 如果该函数能够完成这个任务, 就认为该函数也能够把现有的社交网络矩阵 X 映射到目标矩阵 Y , 也就可以完成链接预测任务. 现有链接可以认为是从完整链接中随机抽取后所形成的, 而完整链接中的任意一个链接被抽取到概率是相同的, 因此可以假设 X 是通过从 Y 进行均匀随机抽取后所形成的矩阵.

对本文函数训练来讲, 除了社交网络本身的链接信息可以被利用外, 用户特征对于函数训练也具有相当的价值, 把这些信息也整合到模型中, 可以提高训练效果. 在训练阶段, 将产生一个“破坏的”矩阵集合 $\tilde{\mathcal{X}}$, 每个 $\tilde{X} \in \tilde{\mathcal{X}}$. 本文将把已知链接矩阵 X 中的 1 以给定的概率 p ($0 < p < 1$) 变为 0, 从而产生一系列“破坏的”矩阵 $\tilde{X}$. 具体通过对以下函数的最小化来训练映射函数 h :

$$\ell(h) = \alpha g(v) + \|Y - h(X)\|_F^2 + \beta \sum_{\tilde{X} \in \tilde{\mathcal{X}}} \|X - h(\tilde{X})\|_F^2, \quad (1)$$

其中 $g(v)$ 是一个衡量用户之间相似性的函数; Y 是希望发现的具有完整链接的矩阵; $h(X)$ 是希望训练的函数, 它可以把已有的社交网络矩阵 X 映射到完整的社交网络矩阵 Y ; $h(\tilde{X})$ 能够把“破坏的”矩阵 $\tilde{X}$ 映射到已知矩阵 X ; $\|\cdot\|_F$ 为加权矩阵 Frobenius 范数 (Frobenius norm), $\|A\|_F^2 = \sum_{i,j} (A_{ij})^2$; α 和 β 是权重, $\alpha + \beta = 1$. 在本文中, α 和 β 将通过试错法获取, 即不断调整 α, β 的值并计算对应的函数值, 最后取能使函数值最优的 α 和 β 值作为参数值 (所有公式说明见附录 A, 表 A1).

3.2 用户之间的相似性函数

每个用户都拥有相当数量的特征, 这些特征可以用来描述用户, 如果用户之间具有相似的特征, 或者共同具有某个特征, 用户之间将有可能建立相互链接. 因此, 基于特征之间的用户相似性的度量对链接的建立是非常重要的. 设 $V_{k \times n}$ 是一个社交网络矩阵, 包含 n 个用户, 每个用户有 k 个特征. 任意两个用户之间相似性度量可以用以下公式表示:

$$g(v) = \sum_{ij} s_{ij} \|v_i - v_j\|_R^2, \quad (2)$$

其中 s_{ij} 是一个常数, 可以通过先验知识获取. 式 (2) 可以被改写为以下矩阵形式:

$$g(v) = \text{tr}(VLV^T), \quad (3)$$

其中 $L = D - S$; S 是社交网络相似性矩阵; 矩阵 D 的对角线元素是矩阵 S 中每一行元素的和, 其他元素为 0; 目标矩阵 $Y = V^T V$, $\text{tr}(\cdot)$ 表示矩阵的迹.

3.3 基于间接链接的矩阵重构

本文希望通过现有的社交网络链接矩阵 X 去重构目标矩阵 Y , 但是并不知道目标矩阵, 因此希望能够训练一个映射函数 $h \in \mathcal{H}$, 把“破坏的”矩阵 $\tilde{X}$ 映射到矩阵 X . 如果能够获取该函数, 就可以完成本文的目标, 通过 X 获取 Y . 因此, 重构的关键在于对映射函数 $h \in \mathcal{H}$ 的获取, 该函数能够产生隐藏在现有用户链接下的内在网络拓扑结构.

两个原本没有链接的用户之间通过第 3 个用户建立相互链接而形成的矩阵用 W 来表示, 该矩阵并不包括原本就存在的链接. 具体来说, 想要知道两个用户 e_i 和 e_j 之间是否可以建立链接, 可以在网络中进行搜索, 看看 e_i 和 e_j 之间是否存在一个用户 e_k , 该用户和 e_i, e_j 之间都存在链接. 如果可以找

到这样的 e_k , 就认为 e_i 和 e_j 也应该存在链接, 否则两者之间就不应该存在链接. 本文定义一个 $n \times n$ 的密切关系矩阵 W , 该矩阵的值表示一个用户和其他用户之间是否存在链接. 可以把该密切关系矩阵 W 和现有的链接矩阵 X 进行线性组合来表示上述通过中间链接而形成的链接, 公式如下所示:

$$Y_{ij} = \sum_k W_{ik} X_{kj}, \quad (4)$$

如果 e_i 和 e_j 之间存在链接, 则 W_{ik} 是正值, 否则是负值.

依据式 (4), 可以通过 X 重构 Y , Y 中所有存在的链接 Y_{ij} 可以表示如下:

$$Y_{ij} = h(X) = \sum_k W_{ik} X_{kj} + b_i, \quad (5)$$

b_i 是一个偏差值. 以矩阵形式, 式 (5) 可以被改写为以下函数形式:

$$(Y)_{ij} = h(X)_{ij} = (WX + b1_n^T)_{ij}, \quad (6)$$

W 是密切关系矩阵, 1_n 是维数为 n 的所有元素都是 1 的列向量.

3.4 低秩近似

可以假定间接亲密关系矩阵包含若干隐含特征, 和很多其他隐变量模型类似, 可以把间接亲密关系矩阵用低秩矩阵来表示: $W = UU^T$. 与 3.1 小节中 α 和 β 取值类似, 合适的维数可以通过试错法获得. 因为减少了冗余特征, 比起原矩阵, 低秩矩阵能够提高预测的精度. 而且低秩矩阵还可以解决过拟合问题. 低秩矩阵表示形式如下:

$$(Y)_{ij} = h(X)_{ij} = \sum_k u_i^T u_k X_{kj} + b_i, \quad (7)$$

u 是矩阵 U 的列向量, 实验中假设矩阵的维数为 $n \times k$, 其中 k 为隐变量的维数, 称为隐维数 (latent dimension). 尽管从上述叙述中可以看出, 亲密关系矩阵 W 是一个对称矩阵, 但并不限定 X 和 Y 也是对称矩阵.

考虑“破坏的”训练矩阵 $\tilde{X}$ 和现有矩阵 X , 模型参数可以通过对以下目标函数进行最小化来获取:

$$\ell_1(V, U, b) = \alpha \text{tr}(VLV^T) + \|V^T V - UU^T X - b1_n^T\|_F^2 + \beta \sum_{\tilde{X} \in \tilde{\mathcal{X}}} \|X - UU^T \tilde{X} - b1_n^T\|_F^2. \quad (8)$$

3.5 梯度计算

与大多数监督学习和去噪模型类似^[33], 使用的训练数据 (“破坏的” 矩阵 $\tilde{X}$) 越多, 得到的结果越好. 但随着训练数据集的增加, 训练时间也将会随着训练集线性增加, 这使得计算并不可行. 受文献 [31] 启发, 依据弱大数定律, 通过求取期望, 可以考虑使用有限的训练集来进行训练, 式 (8) 可以被改写为下述形式:

$$\ell_2(V, U, b) = \alpha \text{tr}(VLV^T) + \|V^T V - UU^T X - b1_n^T\|_F^2 + \beta \mathbb{E} \left[\|X - UU^T \tilde{X} - b1_n^T\|_F^2 \right]_{p(\tilde{X}/X)}, \quad (9)$$

表 1 社交网络特性表

Table 1 The social network properties of the four datasets evaluated

Social network	Nodes	Links	Density
Facebook	2634	63557	0.01833
Twitter	4292	142690	0.01550
Gplus	4712	377848	0.03404
Pokec	12781	80920	0.00099

其中 $p(\tilde{X}/X)$ 是破坏率的 Bernoulli 分布; 期望是关于随机变量 $\tilde{X}$. 本文使用梯度下降法来优化式 (9), 直至目标函数收敛为止, 获取参数值, 从而获取映射函数 h , 通过其获取目标矩阵. 同时需要计算和 V , U , b 相关的梯度, 结果如下:

$$\begin{aligned}\frac{\partial \ell_2}{\partial V} &= \alpha V L + 2 V V^T V - V X^T U U^T - V U U^T X - V \text{blt}, \\ \frac{\partial \ell_2}{\partial U} &= -X V^T V U - V^T V X^T U + P U U^T U + U U^T P U + \text{blt} X^T U \\ &\quad + (1 + \beta) X I b U - 2 \beta P U + \beta Q U U^T U + \beta U U^T Q U + \beta \text{blt} X^T U, \\ \frac{\partial \ell_2}{\partial b} &= -V^T V I + (1 + \beta) U U^T X I + (1 + \beta) \text{blt} I - \beta X I,\end{aligned}$$

其中 $P = X X^T$, $Q = P$, 但 Q 的对角线元素是 P 的对角线元素除以 $(1 - \text{破坏率})$; I 是维数为 n 的列向量, 向量中的值均为 1, $\text{blt} = b I^T$.

算法的复杂度为 $O(kn^2 + nk^2)$, 但因为在运算中采用稀疏矩阵, 实际复杂度更接近于 $O(km + nk^2)$, 其中 m 是稀疏矩阵中非 0 元素的个数. 需要指出, 本文的目标函数并不是凸函数, 因此最终结果可能不是全局最优. 尽管如此, 从实验结果可以看出, 较好的局部最优解也能够带来满意的预测精度.

4 实验分析

4.1 实验设定

本文将和 5 个社交网络链接预测方法进行对比: SpectralLink^[20], CMA-ES^[22], SLP-SAN-VI^[27], TAS+PNR^[31] 以及 mDA^[18]. SpectralLink 算法采用谱聚类方法进行链接预测, 可以计算类内及不同类节点的相似度; CMA-ES 是基于领域相似性和节点相似性的线性组合模型, 并采用自适应协方差矩阵进化策略优化模型权重; SLP-SAN-VI 是对社交-属性网络方法的扩展; TAS+PNR 方法是基于 boosting 学习算法的一种新的迁移学习框架; mDA 是基于矩阵“去噪”思想的链接预测算法. 使用的数据主要来自 4 个社交网络, 所有数据均来自斯坦福大学 Jure Leskovec 教授实验室公开数据: Facebook, Gplus, Twitter, Pokec (斯洛伐克社交网络)^[34]. 其中 Facebook, Gplus, Twitter 为较小规模的数据集, Pokec 为大规模数据集. 这些数据都包含两个部分, 一部分是用户相似矩阵, 另一部分是用户链接矩阵. 表 1 列出了本文实验所用的 4 个网络的特性.

对每一个社交网络数据集, 采用 80% 的数据进行训练, 余下 20% 数据用于测试. 本文对每个方法运行 10 次, 取 AUC (i.e., area under receiver operating characteristic) 的平均值作为最终的运行结果. 本文所用电脑配置如下: 3.60 GHz, i7-4790, 内存为 16.0 GB.

表 2 SJDM 算法参数表
Table 2 Parameters of SJDM algorithm

Social network	α	β	p	k
Facebook	0.1	0.9	0.92	100
Twitter	0.1	0.9	0.92	100
Gplus	0.1	0.9	0.92	100
Pokec	0.1	0.9	0.95	200

表 3 6 个方法在 4 个数据集上的 AUC 值对比
Table 3 The AUC scores of six algorithms tested on four datasets

	Facebook	Gplus	Twitter	Pokec
Spectralink	0.8102 ± 0.0011	0.8014 ± 0.0016	0.8201 ± 0.0014	0.7926 ± 0.0013
CMA-ES	0.8597 ± 0.0009	0.8368 ± 0.0019	0.8586 ± 0.0011	0.8211 ± 0.0017
SLP-SAN-VI	0.8855 ± 0.0030	0.8781 ± 0.0023	0.8913 ± 0.0025	0.8812 ± 0.0015
TAS+PNR	0.9012 ± 0.0017	0.8890 ± 0.0010	0.9088 ± 0.0012	0.8795 ± 0.0024
MDA	0.9189 ± 0.0016	0.9035 ± 0.0011	0.9168 ± 0.0009	0.8966 ± 0.0013
SJDM	0.9564 ± 0.0021	0.9373 ± 0.0003	0.9587 ± 0.0006	0.9250 ± 0.0031

本文方法主要涉及 4 个参数, α 和 β 为权重 (见 3.1 小节), p 为破坏率 (在学习时对现有社交网络的破坏比率), k 为隐维数 (见 3.4 小节), 各个参数值均通过试错法反复实验确定, 其值如表 2 所示.

从表 2 中可知 4 个数据集的 α 和 β 值都相同, 可以看出, 社交网络现有链接的重要性大于用户之间的相似度. 对于 p 和 k , Facebook, Gplus, Twitter 3 个数据集的参数值相同, 而 Pokec 数据集的两个参数值与前 3 个数据集不相同. 分析其原因, 是因为 Pokec 数据集过于稀疏 (见表 1), 对其学习要选用更大的破坏率, 因此其 p 值较大; 而由于其包含 12781 个节点, 远超过其他 3 个社交网络, 因此隐维数 k 的值比其他 3 个数据集要大.

4.2 实验结果

本文在上述 4 个数据集上运行了所有的 6 个方法, 实验结果如表 3 所示. 从表 3 可以看出, 无论对于小规模数据集 Facebook, Twitter, Gplus 还是对大规模数据集 Pokec, SJDM 算法相对于已有算法都较大地提升了 AUC 值. SpectraLink 算法和 CMA-ES 算法属于基于特征的方法, 其预测更多地基于特征, 但社交网络链接结构信息对于链接预测也起着至关重要的作用, 因此其预测精度有所欠缺. SLP-SAN-VI 算法和 TAS+PNR 算法属于隐变量模型, 但这两种算法无法同时获取“同质性”和“统计上相同”这两种特性, 因而其预测精度低于 SJDM 算法. mDA 算法无法利用用户特征信息, 使得其预测精度不足. 而 SJDM 算法不仅可以同时利用特征信息和链接结构信息, 而且能够同时获取“同质性”及“统计上相同”两种特性链接, 因而能够取得比上述 4 个算法更好的实验结果. 实验结果也证明了 SJDM 算法及其隐含假设对链接预测的有效性, 而且 SJDM 算法不但适用于小规模数据集, 对大规模数据集同样有效.

4.3 原矩阵和低秩矩阵实验结果对比

本文采用低秩矩阵 ($n \times k$, $k \ll n$, k 为隐维数) (式 (7)) 代替原矩阵 ($n \times n$) (式 (6)), 低秩矩阵

图 3 低秩矩阵和原矩阵训练时间和 AUC 值

Figure 3 The average AUC score and training time of SJDM in (a) low-rank matrix, (b) original matrix

表 4 低秩矩阵和原矩阵训练时间比

Table 4 Training time ratio of low-rank matrix to original matrix

Social network	Time ratio (Low rank matrix training time / Original matrix training time) (%)
Facebook	5.08
Twitter	3.82
Gplus	3.86
Pokec	0.46

不但可以减少运行时间,而且可以提高精度,同时也可以防止算法过拟合.算法运行时间和矩阵分解之间的关系符合复杂度分析.本文使用原矩阵和低秩矩阵 (Facebook, Gplus, Twitter 的隐维数为 100, Pokec 的隐维数为 200) 在 4 个数据集上进行了验证,其结果如图 3 所示.其中 Pokec 数据集所用电脑配置为 E5-4607v2 (4 处理器),内存 80 GB,其他数据集仍然用 4.1 小节所列配置的电脑.

从图 3 可知低秩矩阵训练所需时间远低于原矩阵训练所需时间,两者时间比见表 4.而且大规模数据集的低秩矩阵所用时间比更是低于小规模数据集,证明了本文算法在大规模数据集上的高效性.从 AUC 值来看,低秩矩阵的 AUC 值也比原矩阵要好,尽管提升的并不多.因此,综合效率和精度来看,低秩矩阵表现出了巨大的优势.

4.4 参数分析

SJDM 算法中最重要的参数有两个:破坏率 p (指现有链接被破坏的比率) 和隐维数 k ,这两个变量的值对于算法结果有着直接而显著的影响.因此,这一部分主要对这两个参数的取值进行了分析,主要分析不同的参数值对于 AUC 值的影响.

首先固定 Pokec 数据集的 p 值为 0.95,其他 3 个数据集 p 值为 0.92.对 Pokec 数据集, k 值取 40 到 200,其他 3 个数据集 k 值取 20 到 100,来评估本文算法对于 k 值的敏感性,4 个数据集的 AUC 平均值如图 4 所示.从图 4 可以看出,随着 k 值的增大, AUC 值也在增大,但变化幅度在逐渐减小.

p 值通常和社交网络矩阵的链接密度相关,对于非常稀疏的社交网络链接矩阵,通常选择较大的

图 4 不同 k 值下的 AUC 值

Figure 4 The AUC score of different latent dimensions k . (a) Facebook, Twitter, Gplus; (b) Pokec

图 5 不同 p 值下的 AUC 值

Figure 5 The AUC score of different corruption p . (a) Facebook, Twitter, Gplus; (b) Pokec

p 值, 较大的值能帮助有效构建“破坏的”训练矩阵. 如果 p 值太小, “破坏的”矩阵将和原矩阵几乎相同, 这将导致无法训练出合适的函数, 也就无法通过该函数获取所需的目标链接矩阵. 固定 Pokec 数据集 k 值为 200, 选取 0.55 到 0.95 之间的 p 值; 固定其他 3 个数据集 k 值为 100, 选取 0.52 到 0.92 之间的 p 值, 并计算对应 p 值下 4 个数据集的 AUC 值, 结果如图 5 所示. 从图 5 可以看出, 随着 p 值的增大, 4 个非常稀疏的数据集的 AUC 值也在随之变大.

本文采用梯度下降法对所构建的优化问题求解, 从而获取所需函数. 在优化过程中, 迭代次数对于优化结果有着直接的影响, 图 6 ~ 9 反映了 4 个数据集不同迭代次数下优化问题的函数值. 从图中可以看出, 函数值成下降趋势, 最后 (迭代到 500 次) 收敛到某一函数值. 前 50 次迭代时函数值下降幅度最大, 后面逐渐下降但趋于平缓, 幅度越来越小, 直至最后收敛. 由于第一次迭代目标函数值太大, 为更好地反映下降趋势, 故把第一次迭代的函数值与其他迭代的函数值分开.

图 6 Facebook 数据集的迭代次数及目标函数值

Figure 6 The objective value with regard to the number of iterations for gradient descent in Facebook

图 7 Twitter 数据集的迭代次数及目标函数值

Figure 7 The objective value with regard to the number of iterations for gradient descent in Twitter

4.5 实例验证: 链接预测及社区发现

为了更好地验证本文算法,同时也体现该算法在实际中的应用价值,本文从所在大学中选取 180 个学生构成一个社交网络矩阵,命名为 xuesheng180,采用本文方法对其进行研究. xuesheng180 矩阵中,全部链接应该是 16110,现有的相互认识的链接为 2134 个. 180 名学生或因住同宿舍、同班级,或因某些兴趣爱好(比如打篮球等)等构成了 10 个小团体(即社区).

为了验证本文算法的有效性,本文随机抽取了 xuesheng180 中 2134 个链接中的 1100 个链接,用 SJDM 算法对抽取后的矩阵进行了链接预测,本文参数设置如 4.1 小节所示. 实验结果表明,1100 个链接中的 1062 个链接被成功预测出,准确率达到 0.9655. 对包含 2134 个链接的 xuesheng180 矩阵采用本文方法进行了链接预测,预测后的矩阵命名为 xueshengall180. 采用 ClusterONE (参数 d 值为 0.3)^[35] 方法对原矩阵 xuesheng180 和预测后的矩阵 xueshengall180 进行社区发现. 采用 NMI 指标对社区发现效果进行评价. 采用原矩阵 xuesheng180 和预测后矩阵 xueshengall180 进行社区发现的 NMI 值分

图 8 Gplus 数据集的迭代次数及目标函数值

Figure 8 The objective value with regard to the number of iterations for gradient descent in Gplus

图 9 Pokec 数据集的迭代次数及目标函数值

Figure 9 The objective value with regard to the number of iterations for gradient descent in Pokec

别为 0.832 和 0.943.

该实例也再次验证了 SJDM 算法预测的有效性. 而使用预测后的权重矩阵进行社区发现的 NMI 值比采用原矩阵的 NMI 值有了较大提升, 这也证明了本文算法的价值. 在实际中, 可以采用本文算法对学生的相互联系和兴趣社区进行预测和发现, 这将可以促进学生在日常生活中的交友及兴趣活动的推广, 对学生的管理将是个很好的促进.

5 结论与展望

社交网络在人们生活中越来越重要, 本文对社交网络中的重要问题——链接预测进行了研究. 本文把链接预测问题看作一个有监督矩阵“去噪”问题. 本文主要是构建一个映射函数, 该映射函数能够

把“破坏的”矩阵(相对于已知矩阵)映射到已知矩阵,也意味着该函数能够把已知矩阵映射到充满所有链接的目标矩阵,从而完成链接预测.而采用预测后的权重矩阵进行社区发现的精度也比采用原矩阵要高.实验结果也验证了 SJDM 算法的有效性.但在本文算法中,对于两个社交网络用户之间是否可以建立链接,主要考虑的一个因素是两个用户之间是否有共同朋友(是 local 的),而没有考虑两个用户的朋友之间是否具有朋友关系(是 global 的),这种关系也可以帮助建立两个用户的联系.比如用户 u_1 有一个朋友 u_2 , u_4 有一个朋友 u_3 , 如果 u_2 和 u_3 是朋友,那么 u_1 和 u_4 也可能成为朋友.这种搜索是基于全局的,对于链接预测是有价值的,这将是下一步研究的一个方向.

参考文献

- 1 Wang P, Xu B W, Wu Y R, et al. Link prediction in social networks: the state-of-the-art. *Sci China Inf Sci*, 2015, 58: 011101
- 2 Aiello L M, Barrat A, Schifanella R, et al. Friendship prediction and homophily in social media. *ACM Trans Web*, 2012, 6: 1–33
- 3 Mori J, Kajikawa Y, Kashima H, et al. Machine learning approach for finding business partners and building reciprocal relationships. *Expert Syst Appl*, 2012, 39: 10402–10407
- 4 Wu S, Sun J, Tang J. Patent partner recommendation in enterprise social networks. In: *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM'13)*, Rome, 2013. 43–52
- 5 Tang J, Wu S, Sun J M, et al. Cross-domain collaboration recommendation. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*, Beijing, 2012. 1285–1293
- 6 Pavlov M, Ichise R. Finding experts by link prediction in co-authorship networks. In: *Proceedings of the 2nd International ISWC+ASWC Workshop on Finding Experts on the Web with Semantics (FEWS)*, Busan, 2007. 42–55
- 7 Wohlfarth T, Ichise R. Semantic and event-based approach for link prediction. In: *Proceedings of the 7th International Conference on Practical Aspects of Knowledge Management (PAKM'08)*, Yokohama, 2008. 50–61
- 8 Bhattacharyya P, Garg A, Wu F S. Analysis of user keyword similarity in online social networks. *Soc Netw Anal Min*, 2011, 1: 143–158
- 9 Akcora C G, Carminati B, Ferrari E. User similarities on social networks. *Soc Netw Anal Min*, 2013, 3: 475–495
- 10 Anderson A, Huttenlocher D, Kleinberg J, et al. Effects of user similarity in social media. In: *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM'12)*, Seattle, 2012. 703–712
- 11 Adamic L A, Adar E. Friend and neighbors on the web. *Soc Netw*, 2003, 25: 211–230
- 12 Lichtenwalter R N, Chawla N V. Vertex collocation profiles: subgraph counting for link analysis and prediction. In: *Proceedings of the 21st World Wide Web Conference (WWW'12)*, Lyon, 2012. 1019–1028
- 13 Symeonidis P, Mantas N. Spectral clustering for link prediction in social networks with positive and negative links. *Soc Netw Anal Min*, 2013, 3: 1433–1447
- 14 Valverde-Rebaza J, Lopes A A. Exploiting behaviors of communities of twitter users for link prediction. *Soc Netw Anal Min*, 2013, 3: 1063–1074
- 15 Liu H, Hu Z, Haddadi H, et al. Hidden link prediction based on node centrality and weak ties. *Europhys Lett*, 2013, 101: 18004
- 16 Zhu J. Max-margin nonparametric latent feature models for link prediction. In: *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, 2012. 1179–1186
- 17 Lloyd J R, Orbanz P, Ghahramani Z, et al. Random function priors for exchangeable arrays with applications to graphs and relational data. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Lake Tahoe, 2012. 1007–1015
- 18 Chen M M, Xu Z X, Weinberger K Q, et al. Marginalized denoising autoencoders for domain adaptation. In: *Proceedings of International Conference on Machine Learning 2012 (ICML2012)*, Edinburgh, 2012
- 19 Papadimitriou A, Symeonidis P, Manolopoulos Y. Fast and accurate link prediction in social networking systems. *J Syst Softw*, 2012, 85: 2119–2132
- 20 Symeonidis P, Iakovidou N, Mantas N, et al. From biological to social networks: link prediction based on multi-way spectral clustering. *Data Knowl Eng*, 2013, 87: 226–242

21 Bastani S, Jafarabad A K, Zarandi M H F. Fuzzy models for link prediction in social networks. *Int J Intel Syst*, 2013, 28: 768–786

22 Bliss C A, Frank M R, Danforth C M, et al. An evolutionary algorithm approach to link prediction in dynamicsocial networks. *J Comput Sci*, 2014, 5: 750–764

23 Bayrak A E, Polat F. Contextual feature analysis to improve link prediction for location based social networks. In: *Proceedings of the 8th Workshop on Social Network Mining and Analysis (SNAKDD'14)*, New York, 2014

24 Xie X Q, Li Y J, Zhang Z Q, et al. A joint link prediction method for social network. In: *Proceedings of Intelligent Computation in Big Data Era*, Harbin, 2015. 56–64

25 Zhu J, Xie Q, Chin E J. A hybrid time-series link prediction framework for large social network. In: *Proceedings of the 23rd Database and Expert Systems Applications (DEXA2012)*, Vienna, 2012. 345–359

26 Liu F, Liu B Q, Sun C J, et al. Deep learning approaches for link prediction in social network services. In: *Proceedings of the 20th International Conference on Neural Information Processing (ICONIP2013)*, Daegu, 2013. 425–432

27 Gong N Z, Talwalkar A, Mackey L, et al. Joint link prediction and attribute inference using a social-attribute network. *ACM Trans Intel Syst Technol*, 2014, 5: 529–544

28 Panwar A P S, Niyogi R. A heuristic for link prediction in online social network. In: *Intelligent Distributed Computing*. Berlin: Springer, 2015. 31–41

29 He Y L, Liu J N K, Hu Y X, et al. OWA operator based link prediction ensemble for social network. *Expert Syst Appl*, 2015, 42: 21–50

30 Sherkat E, Rahgozar M, Asadpour M. Structural link prediction based on ant colony approach in social networks. *Phys A*, 2015, 419: 80–94

31 Nguyen-Thi A T, Nguyen P Q, Ngo T D, et al. Transfer adaBoost SVM for link prediction in newly signed social networks using explicit and PNR features. *Procedia Comput Sci*, 2015, 60: 332–341

32 Hoff P D. Modeling homophily and stochastic equivalence in symmetric relational data. *arXiv:0711.1146*

33 Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders. In: *Proceedings of the 25th International Conference on Machine learning*, Helsinki, 2008. 1096–1103

34 McAuley J, Leskovec J. Learning to discover social circles in ego networks. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Lake Tahoe, 2012. 539–547

35 Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Meth*, 2012, 9: 471–472

附录 A 公式说明表

表 A1 公式说明表
Table A1 Formula description table

Formula number	Formula description
(1)	The initial function of the objective function
(2)	User similarity function
(3)	The matrix representation of formula (2)
(4)	Link formation based on close relation matrix
(5)	Matrix reconstruction based on formula (4)
(6)	The matrix representation of formula (5)
(7)	Matrix low rank approximation
(8)	The concrete objective function
(9)	Transformation of objective functions based on weak law of large numbers

Link prediction in online social networks based on supervised joint denoising model

Zhangang HAO^{1*}, Weixiong ZHANG² & Zheng CHEN²

1. *School of Business Administration, Shandong Technology and Business University, Yantai 264005, China;*

2. *Department of Computer Science and Engineering, Washington University, St. Louis 63130, USA*

* Corresponding author. E-mail: zghao2000@sina.com

Abstract Link prediction of social networks can capture important information about missing links for applications in many fields. Because of the failure to make full use of information as well as capture all properties, the link prediction precision of most of methods is low. For higher precision, we propose a novel algorithm, a supervised joint denoising model (SJDM) that formulates the link prediction problem as a supervised matrix “denoising” problem. The central piece of our method is a function that is trained using the features of users and topological structures of social networks. The function can map the observed “corrupted” matrix to an “uncorrupted” matrix (target matrix). We performed community detection using the target matrix, which is better than using the original matrix. Five real networks are processed with this algorithm. The results show that SJDM algorithm is more efficient compared to the other four algorithms.

Keywords social network analysis, prediction, supervised learning, community networks, matrix “denoising”, weak law of large number

Zhangang HAO was born in 1976. He received his Ph.D degree in management science and engineering from Tianjin University, Tianjin, in 2006. He is currently an associate professor at Shandong Technology and Business University. His research interests include social networks and deep learning.

Weixiong ZHANG was born in 1960. He received his Ph.D degree from the University of California-Los Angeles, Los Angeles, in 1994. He is currently a professor at Washington University in St. Louis. His research interests include artificial intelligence and computational biology.