DOI: 10.11992/tis.201710027
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180404.1358.012.html

符号网络的局部标注特征与预测方法

苏晓萍[1], 宋玉蓉[2]
（1. 南京工业职业技术学院 计算机与软件学院, 江苏 南京 210046; 2. 南京邮电大学 自动化学院, 江苏 南京 210003）

摘　要: 当复杂网络的边具有正、负属性时称为符号网络。符号为正表示两用户间具有相互信任 (朋友) 关系, 相反, 符号为负表示不信任 (敌对) 关系。符号网络中的一个重要研究任务是给定部分观测的符号网络, 预测未知符号。分析发现, 具有弱结构平衡特征的符号网络, 其邻接矩阵呈现全局低秩性, 在该特征下链路符号预测问题可以近似表达为低秩矩阵分解问题。但基本低秩模型中, 相邻节点间符号标注的局部行为特征未得到充分利用, 论文提出了一种带偏置的低秩矩阵分解模型, 将邻居节点的出边和入边符号特征作为偏置信息引入模型, 以提高符号预测的精度。利用真实符号网络数据进行的实验证明, 所提模型能够获得较其他基准算法好的预测效果且算法效率高。

关键词: 符号网络; 符号预测; 低秩; 矩阵分解; 标注偏置; 结构平衡理论; 弱结构平衡理论; 地位理论
中图分类号: TP399　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)03−0437−08

中文引用格式: 苏晓萍, 宋玉蓉. 符号网络的局部标注特征与预测方法[J]. 智能系统学报, 2018, 13(3): 437–444.
英文引用格式: SU Xiaoping, SONG Yurong. Local labeling features and a prediction method for a signed network[J]. CAAI transactions on intelligent systems, 2018, 13(3): 437–444.

Local labeling features and a prediction method for a signed network

SU Xiaoping[1], SONG Yurong[2]
(1. School of Computer and Software Engineering, Nanjing Institute of Industry Technology, Nanjing 210046, China; 2. College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: A complex network may be considered as a symbol network when links have a positive or negative sign attribute. In signed social networks, positive links represent a trust (friends) relationship between users. In contrast, negative links indicate distrust (hostility). An important task in a signed network is to define a signed network based on partial observation to predict an unknown symbol. Through analysis, we found that for a signed network with weak structural balance, its adjacent matrix has a global low-rank characteristic and the prediction of the link sign can be approximated as a low-rank matrix factorization. However, in a basic low-rank model, it is difficult to sufficiently utilize the local behavior features for labeling the signs of links between the neighboring nodes. Herein, a low-rank matrix factorization model with bias was proposed. In this model, the sign features of the exit and entry links of a neighboring node were introduced to improve the precision of sign prediction. Experiments based on real data revealed that the low-rank model with bias can obtain better prediction results than other benchmark algorithms and that the proposed algorithm performed with a high efficiency.

Keywords: signed networks; sign prediction; low rank; matrix factorization; signed bias; structural balance theory; weak structural balance theory; status theory

符号网络是指边具有正或负符号属性的网络, 符号为正表示网络中两节点间具有相互信任的、积极的朋友关系, 负边则表示不信任的、消极的敌对关系。具有符号属性的网络普遍存在[1], 研究链路

收稿日期: 2017−10−30.　**网络出版日期:** 2018−04−04.
基金项目: 国家自然科学基金项目 (61672298, 61373136); 教育部人文社会科学研究规划基金项目 (17YJAZH071); 江苏省高校优秀科技创新团队项目.
通信作者: 苏晓萍. E-mail: 419033424@qq.com.

的符号属性有利于理解网络的基本结构特征，理解信任和不信任的传播方式。另外，社会符号网络的边符号属性能够直接反映节点间的态度，因此在推荐系统[2]、舆情分析与观点形成[3]、网络欺凌与社会排斥[4]等问题中均能通过符号分析获得应用。符号网络的研究始于 Heider[5]基于社会心理学对人类关系的研究，而随着复杂网络研究兴起，符号网络的结构特征与演化规律受到更多研究者的关注[6-7]，如何通过部分观测到的网络符号预测未知的边符号成为符号网络中非常重要的研究方向。

符号预测方法大致可以分为两类：1) 考虑网络局部特征的方法；2) 考虑网络全局特征的方法。考虑网络局部特征的方法主要利用节点的邻域特征，如：节点出边、入边的符号以及相邻三元组各边标注符号特征进行符号预测。这类方法主要基于网络局部特征以及社会学相关理论实现边符号的预测，所有基于弱结构平衡[8]和地位理论[9]的预测方法均要求两节点间具有共同邻居时算法才有效，但统计结果发现，现实的符号网络中有很大比例的节点不能构成三元组。Guha 等[10]最早基于网络模型研究符号预测问题，他们将信任网络表示为矩阵并运用不同的矩阵运算代表信任关系在网络上的不同传播方式，实现了信任关系的预测。Leskovec 等[11]采用机器学习的方法对符号预测问题进行了研究，他们利用节点的出度、入度、节点的嵌入性以及基于地位理论的所有 16 种待预测边所处的三角形的关系模式作为特征采用逻辑回归模型训练分类器，得到了较高的预测精度。文献[12]则通过网络局部特征和地位理论为特征采用 SVM 算法进行二值分类实现符号预测。相对于 Leskovec 考虑长度为 3 的有序环构建的网络特征，Chiang 等[13]利用 Katz 指标提出一个不平衡测度指标并通过长度为κ的环的平衡程度构建特征集，然后使用逻辑回归模型进行符号预测，当环的长度从 3 增加到 5 时，预测精度有所提高，但是当$\kappa>5$后对预测精确度的影响不大。文献[14]指出：能够反映符号网络不平衡程度的测度都可以用于符号预测。文献[15]通过分析两节点间不同的连接形式，提出符号预测的方法，使得在没有共同邻居的情形下的预测精度有所提高；符合符号网络局部倾向于结构平衡或弱结构平衡的特征反过来会促使符号网络的全局特征出现，因此有很多利用网络全局结构进行符号预测的方法。文献[16]就从谱分析的角度出发进行符号预测，并指出许多基于谱分析的方法可以从简单的二值网络扩展到符号网络。他们将拉普拉斯矩阵的定义扩充到符号网络，通过拉普拉斯矩阵的核函数进行网络符号的预测。Hsieh[17]等发现满足弱结构平衡理论的符号网络其邻接矩阵具有低秩特征，于是将符号预测问题转化为矩阵填充问题，用低秩填充法有效地进行了符号预测。他们还将符号预测近似为低秩矩阵分解问题进行了符号预测。文献[18]也研究了矩阵分解在符号预测中的应用并解决了数据不平衡对预测精度的影响。文献[19]提出了一种区别于 Hsieh 以逐点误差衡量原矩阵与结果矩阵误差的方法，他们将成对误差应用到矩阵分解的损失函数中，给出的算法 MF-LiSP 取得了较高精确度。通过以上介绍发现，符号网络的局部结构特征与全局特征联系紧密，符号预测方法仅使用局部特征或全局特征都不够全面，在预测算法中如何同时利用局部和全局特征是一个值得研究的问题。

受以上研究的启发，从真实网络数据的统计分析出发，结合节点局部标注特征和网络全局结构特征设计了一种新的基于低秩矩阵分解的符号预测模型，解决了符号网由于数据稀疏和网络局部特征利用不足带来的预测精度不高的问题。

1　相关理论

1.1　基本定义

定义符号网络G为$G=(V,E,S)$，其中$V=\{1,2,3,\cdots,n\}$为节点集合，$E=\{1,2,3,\cdots,m\}$为边集合，$S=\{-1,0,1\}$表示边符号，$i,j\in V$,$e(i,j)\in E$，$s(i,j)\in S$，设O为被观测到的边集，则$O\subseteq E$。符号网络G对应邻接矩阵$\boldsymbol{A}$，其中：

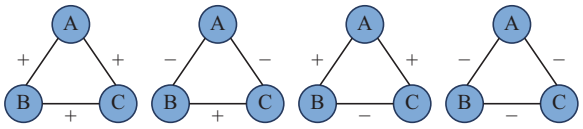
$$A_{ij}=\begin{cases}1, & i与j之间有正边\\ 0, & 边符号未知或不存在边\\ -1, & i与j之间有负边\end{cases}$$

符号网络 G 可能为有向图也可能为无向图，当 G 为有向图时 $\boldsymbol{A}$ 为非对称矩阵，若 G 为无向图则 $\boldsymbol{A}$ 为对称矩阵。

1.2　结构平衡与弱结构平衡理论

结构平衡理论把人与人之间的关系分为积极和消极两种，被形式化地描述为符号网络，边的正、负符号分别表示积极关系和消极关系。此时符号网络中 3 个节点间的关系共形成 4 种三角形模体，如图 1 所示。从社会心理学角度看，以下结论成立：朋友的朋友是朋友；朋友的敌人是敌人。据此判定图 1(a)、(b) 是平衡的，而图 1(c)、(d) 是不平衡的。不平衡的结构具有向平衡结构转换的趋势。在三角形模体中判断局部平衡性时可以通过三条边符号之积来实现：三符号积为正则平衡，否则不平衡。Cartwright 和 Harary[20]将 Heider[2]的社会学结论形式化地描述为图结构并证明符号网络平衡的充分必

要条件是: 网络中的节点能够被划分为两个子集, 每个子集内的所有边均为正, 子集间的边均为负。



(a) 关系模式1 (b) 关系模式2 (c) 关系模式3 (d) 关系模式4

图 1　符号网中 4 种三角形关系模式

Fig. 1　Four relationship patterns of triads in signed network

平衡网络的判别条件比较严苛, 现实中很难找到平衡网络的情形, 因此 Davis[8]放宽结构平衡的约束提出了弱结构平衡理论, 弱结构平衡理论规定: 只要三角形模体中不存在两正一负的关系就构成弱平衡, 在该条件下, 图 1(a)、(b)、(d) 代表的情形均可看作平衡的结构。当网络满足弱平衡结构时, 节点可以被分成κ个子集, 且子集内节点间的边全为正, 子集间节点的边全为负。这类符号网也被称为κ-平衡网。

1.3　矩阵的秩与结构平衡

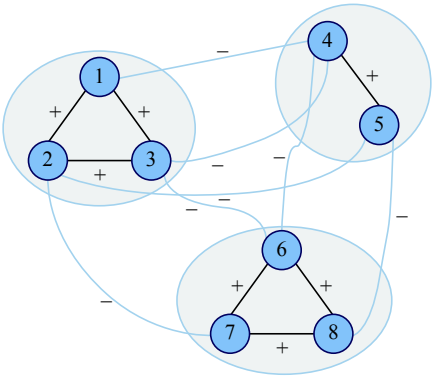
根据弱结构平衡的定义, 当符号网络满足κ-平衡条件时, 网络节点可以被分成κ个子集, 当对网络节点按编号排序, 邻接矩阵将是块对角矩阵。

图 2(a) 给出了一个满足弱平衡网的示例, 图中 8 个节点被分成 3 个子集, 图 2(b) 为图 2(a) 对应的邻接矩阵 A, 若补齐图 2(a) 中缺失的边, 使其成为完全图, 该图对应的邻接矩阵为块对角矩阵 $\boldsymbol{X}$, 块内很明显矩阵的秩Rank($\boldsymbol{X}$) = 3小于矩阵的行列数 8。根据以上分析可知符号网络添加相应具有固定符号的边将使符号网络向完全κ-平衡网靠近。

因此可以把符号预测问题看作矩阵填充问题: 已知被部分观测到的矩阵$\boldsymbol{A}$, 采用矩阵填充的方法找到矩阵$\boldsymbol{X}$。此时符号预测问题可被看作优化问题: 填充矩阵中零值使得目标矩阵$\boldsymbol{X}$的秩最小。该问题可形式化描述为

$$\begin{aligned} &\min \quad \mathrm{Rank}(\boldsymbol{X}) \\ &\text{s.t.} \quad X_{ij}=A_{ij}, \forall e(i,j)\in O, \\ &\qquad X_{ij}\in\{\pm 1\}, \forall e(i,j)\notin O \end{aligned} \tag{1}$$

式 (1) 的目标函数是$\boldsymbol{X}$矩阵的秩, 即其奇异值构成向量的稀疏性, 通常上述优化问题是 NP 难的, 而函数Rank($\boldsymbol{X}$)在矩阵谱范数单位球上的凸包络是矩阵的核范数 (即矩阵所有奇异值的和), 因此可以用凸的核范数最小化来近似秩最小化问题, 文献[21]表明: 当被观测矩阵均匀抽样且$|O| \geqslant C\mu^4 n(\log_2 n)^2$成立时, 矩阵 $\boldsymbol{X}$ 可以以$1-n^{-3}$的概率被恢复。但是对于符号网络来说, 均匀抽样不容易做到, 因为通过 4.1 节数据描述可以看到符号网络 80% 的边为正, 同时矩阵填充的运算速度较慢, 因此矩阵填充也经常用低秩矩阵分解来近似。



(a) 符号网络示例

$$\boldsymbol{A}=\begin{pmatrix} 0 & 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & -1 & 0 & -1 & 0 \\ 1 & 1 & 0 & -1 & 0 & -1 & 0 & 0 \\ -1 & 0 & -1 & 0 & 1 & -1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & -1 & 0 & 0 & 1 & 1 \\ 0 & -1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 1 & 1 & 0 \end{pmatrix}$$

(b) 图2(a)对应的邻接矩阵

$$\boldsymbol{X}=\begin{pmatrix} 0 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 0 & 1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 0 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & 0 & 1 & -1 & -1 & -1 \\ -1 & -1 & -1 & 1 & 0 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & 0 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & -1 & -1 & 1 & 1 & 0 \end{pmatrix}$$

(c) 图2(a)对应的完全图邻接矩阵

图 2　弱平衡结构与矩阵的低秩特性

Fig. 2　Weakly balanced network and low-rank structure

2　符号预测

当符号预测被近似为低秩矩阵分解问题时, 邻接矩阵$\boldsymbol{A}$可被分解为两个κ行n列的矩阵 $\boldsymbol{P}^{\mathrm{T}}$ 和 $\boldsymbol{Q}$, 优化目标是使$\boldsymbol{P}^{\mathrm{T}}\boldsymbol{Q}$与 $\boldsymbol{A}$ 之间的误差最小, 原矩阵可以被$\hat{r}_{ij}=(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{Q})_{ij}$值填充, $\hat{r}_{ij}$就是预测到的用户i对用户j的评价。矩阵分解模型可被形式化地描述为

$$\min_{\boldsymbol{P}^{\mathrm{T}},\boldsymbol{Q}\in\mathbb{R}^{k\times n}} C=\sum_{e(i,j)\in O} l(A_{ij},\sum_{k=1}^{\kappa}P_{ik}Q_{kj})+\lambda\|\boldsymbol{P}\|^2+\lambda\|\boldsymbol{Q}\|^2 \tag{2}$$

式 (2) 中 l 为损失函数用于评价预测值与原矩阵间的误差, 后两项为正则化项, 用来防止过拟合损失函数, 可根据具体问题进行选择。虽然上式是非凸的, 但是实践证明该方法在很多矩阵填充问题

中均取得了很好的预测效果。

基本矩阵分解模型充分利用了邻接矩阵的全局低秩特性，但是，在被符号网络所代表的社会关系网中，不同节点的标注行为常常具有偏置现象：网络“喷子”也被称为“Troll”的节点，该类节点为引起别人的注意会故意攻击其他人，“Troll”节点会发出比其余节点更多的负边；与此相对应的，有些节点会收到低于平均水平的评价，它们可能受到“网络欺凌”，这一现象的社会心理学根源是“认知失调”，人们通常为保持与他人态度的一致而调整自己的行为因此而攻击收到过负面评价的人。从真实符号网络的统计特征发现，这两类节点在符号网中确实存在，虽然数量不多但其作用巨大。在符号预测问题中仅考虑平均后的全局特征并不能完全反映网络结构特征，节点的局部标注特征需要在预测模型中得以体现。现定义待预测边的局部标注特征为

$$b_{ij}=\mu+Ui_{\text{out}}+Uj_{\text{in}} \tag{3}$$

式中：μ为符号网络的平均标注倾向，当μ为负时说明网络用户更倾向于给其他用户以负面评价；μ为正时，则表示网络用户有给其他邻居以正面评价的倾向。设待预测边$e(i,j)$两端的节点为 i 和 j，Ui_{out}表示节点 i 发出的边符号的均值，Ui_{out}的值能够反映节点 i 对相邻节点的局部标注特征：若节点发出的负边数大于正边数，表示节点 i 给邻居以负面评价的可能性大，$e(i,j)$被预测为负的可能性就增加。同理，Uj_{in}为j收到的边符号的均值，当Uj_{in}为负时表示节点j收到了更多的负面评价，因此$e(i,j)$被预测为负的可能性就增加。图 3 给出了符号网络标注的局部偏置示例，设$\mu=0.2$，即符号网络全局有正面评价的倾向，经计算可得：$Ui_{\text{out}}=[(-1)+(-1)+1]/4=-1/4$，$Uj_{\text{in}}=-1/4$，于是$b_{ij}=-3/10$，此时边$e(i,j)$的符号预测结果将向负偏斜。

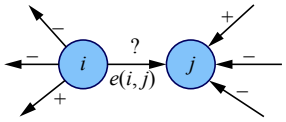


图 3　标注行为的偏置现象

Fig. 3　Bias behavior of signed edges

b_{ij}的值能够很好地反映待预测边两端节点的局部标注行为和行为偏好，将标注偏好反映在预测的目标函数，得到较基本模型更为精细的预测模型：

$$\min_{\boldsymbol{P}^{\mathrm{T}},\boldsymbol{Q}\in\mathbb{R}^{k\times n}} C=\sum_{e(i,j)\in\mathrm{O}}l(A_{ij}-(b_{ij}+\sum_{k=1}^{\kappa}P_{ik}Q_{kj}))+\lambda(\|\boldsymbol{P}\|_1^2+\|\boldsymbol{Q}\|_1^2+Ui_{\text{out}}{}^2+Uj_{\text{in}}{}^2) \tag{4}$$

根据式 (4) 可知：节点 i 对节点 j 的符号可被预测为$\hat{r}_{ij}=b_{ij}+(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{Q})_{ij}$，而式 (2) 表示的基本矩阵分解模型得到的符号预测结果为$\hat{r}_{ij}=(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{Q})_{ij}$。因此，以式 (4) 为目标函数的优化方法，不但考虑了符号网的全局低秩特性还考虑了待预测边两端节点的局部标注特征，与基本模型相似，添加了关于局部特征项的正则化项$Ui_{\text{out}}{}^2+Uj_{\text{in}}{}^2$防止过拟合。

损失函数 l 可以有多种选择，本文选择 Square_loss 为损失函数，于是优化目标函数可写成：

$$\min_{\boldsymbol{P}^{\mathrm{T}},\boldsymbol{Q}\in\mathbb{R}^{k\times n}} C=\sum_{e(i,j)\in\mathrm{O}}\left\|A_{ij}-(b_{ij}+\sum_{k=1}^{\kappa}P_{ik}Q_{kj})\right\|^2+\lambda(\|\boldsymbol{P}\|_1^2+\|\boldsymbol{Q}\|_1^2+Ui_{\text{out}}{}^2+Uj_{\text{in}}{}^2) \tag{5}$$

对式 (5) 给出的优化问题可以采用随机梯度下降法进行求解，令$e_{ij}=A_{ij}-(b_{ij}+\sum_{k=1}^{\kappa}P_{ik}Q_{kj})$，通过求梯度以确定优化函数下降方向：$\dfrac{\partial C}{\partial P_i}=-2e_{ij}Q_j+2\lambda P_i$，$\dfrac{\partial C}{\partial Q_j}=-2e_{ij}P_i+2\lambda Q_j$，同理也可求得目标函数对$Ui_{\text{out}}$、$Uj_{\text{in}}$的偏导数，由于沿梯度方向相反的方向下降最快，于是得到如下迭代公式：

$$P_i\leftarrow P_i+\alpha(e_{ij}Q_j-\lambda P_i) \tag{6}$$

$$Q_j\leftarrow Q_j+\alpha(e_{ij}P_i-\lambda Q_j) \tag{7}$$

$$Ui_{\text{out}}\leftarrow Ui_{\text{out}}+\alpha(e_{ij}-\lambda Ui_{\text{out}}) \tag{8}$$

$$Uj_{\text{in}}\leftarrow Uj_{\text{in}}+\alpha(e_{ij}-\lambda Uj_{\text{in}}) \tag{9}$$

通过反复迭代并不断优化参数，使观测矩阵 $\boldsymbol{A}$ 与分解后矩阵$\boldsymbol{B}+\boldsymbol{P}^{\mathrm{T}}\boldsymbol{Q}$间的误差小于设定的误差值即最终收敛。其中$\alpha$为学习速度，$\alpha$越大下降就越快。随机梯度下降的时间复杂度为$O(tm\kappa)$，$t$为迭代并收敛次数，$m$为节点个数，$\kappa$为秩数。由于符号网络满足低秩特性，通常$\kappa$值很小，且收敛较快，因此采用随机梯度下降法求解最小化问题速度较快。

3　实验结果与分析

3.1　数据集描述

实验中的 3 个真实大型社会网络数据来自于斯坦福大学的 SNAP[2] 项目，Epinions 给出了用户间“who-trust-whom”的关系，Slashdot 是一个技术相关的新闻网站，允许用户根据自身观点标记其他用户为 friend/foe，Wikipedia 是维基百科申请管理员身份的投票关系网，若一个用户被大多数其他用户同意则当选为某一学科的管理员负责百科词条的维护，若该用户未受到大多数其他用户的赞成票则选举失败。表 1 给出了 3 个网络的统计特征。

表 1 的统计结果显示：3 个符号网络中正边占比均在 75% 以上，而负边占比较少，互惠边 (reciprocal edges) 是指两用户间持有相同态度，这样的互惠边在网络中占有一定比例，且互惠边中正边居多，这与人们社会心理有关，当一个人讨厌另一个

人反应的是不予理睬, “爱的反义词不是恨而是冷漠”。而一个人对另一个人的示好通常显示出“镜子效应”。统计结果还发现: 发出 50% 以上负边的节点只占网络节点极少部分, 且大部分的负边都是由一些特定节点发出的, 这些节点充满反社会特征, 并通过攻击别人博取别人的关注, 这些节点发出的新边为负的可能性极大, 而收到 50% 以上负边的节点也的确存在, 即“网络欺凌”是事实, 存在“人云亦云”的现象。统计结果还显示, 符号网络中有一定比例的节点无法构成三元组, 此时基于结构平衡理论的预测算法将失效。

表 1　数据集统计特征

Table 1　Statistics of datasets

统计特征	数据集		
	Epinions	Slashdot	Wikipedia
节点数	131 828	82 144	7 118
边总数	841 372	549 202	104 357
正边占比/%	85.3	77.4	78.4
负边占比/%	14.7	22.6	21.6
互惠正边占比/%	30.2	17.0	5.1
互惠负边占比/%	0.30	0.31	0.28
发出 50% 以上负边占比/%	7.9	6.7	16.9
收到 50% 以上负边占比/%	15.1	19.0	12.2
非三元组边占比/%	20.2	44.8	8.1

3.2　预测效果与分析

为证明所提带有偏置的矩阵分解模型 MF-Bias (matrix factorization with bias) 对符号预测问题的有效性, 将它与以下基准预测算法进行比较。

1) OutDegree(简写为 OD): 若$d^+_{\text{out}}(i) \geqslant d^-_{\text{out}}$, 则被预测边$e(i,j)$的符号为正, 反之为负。

2) InDegree(简写为 ID): 若$d^+_{\text{in}}(j) \geqslant d^-_{\text{in}}(j)$, 则被预测边$e(i,j)$的符号为正, 反之为负。

3) LR (logistic regression)[11]: 将符号预测问题看作二值分类问题, 采用逻辑回归模型训练分类器, 得到了较高的预测精度。

4) MF(matrix factorization)[17]: 由 Hsieh 等提出的基本矩阵分解模型。

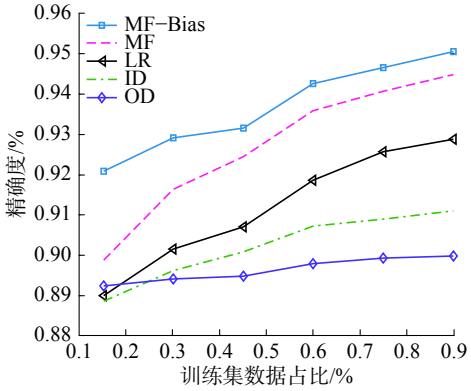
3.2.1　评价指标

1) 均方根误差 (RMSE)

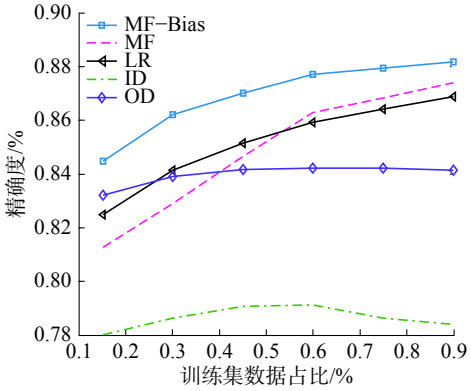
它是衡量模型误差率的常用方法, 反映了观测值与真值偏差的平方和观测次数 n 比值的平方根, 计算公式为

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}} \tag{10}$$

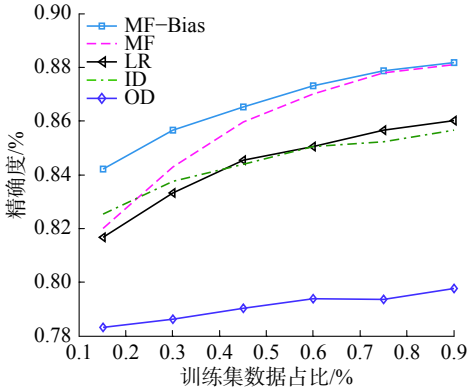
式中: p_i为第i个观测值, a_i为第i个真实值, RMSE 值越小预测误差越小。图 4(a)、(b)、(c) 给出了 3 个符号网络在不同抽样比率下的 RMSE 值。



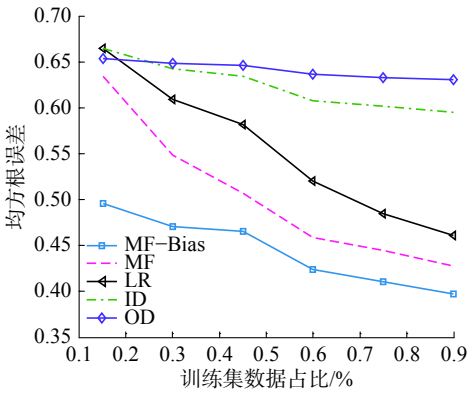
(a) 数据集Epinions的预测精度



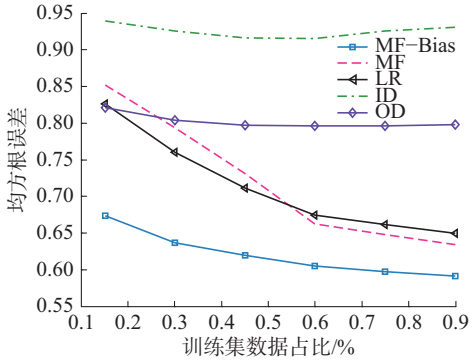
(b) 数据集Slashdo的预测精度



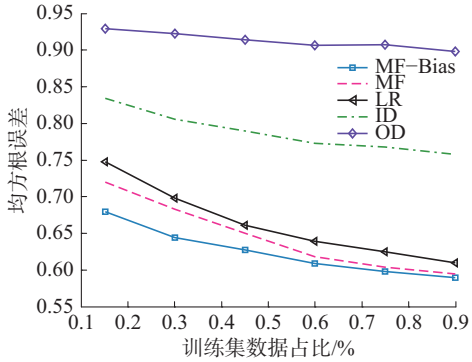
(c) 数据集Wikipedia的预测精度



(d) 数据集Epinions的预测误差



(e) 数据集Slashdot的预测误差



(f) 数据集Wikipedia的预测误差

图 4　3 个符号网络的预测结果

Fig. 4　Three signed networks predicted results

2) 精确性 (Accuracy)

用于评价预测算法对符号预测的准确程度，精确性计算公式为

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{P+N} \tag{11}$$

式中：TP表示对符号为正的边的预测正确的数目，TN表示符号为负的预测正确的数目，$P+N$则是需要预测的边的总数。Accuracy 的值越大表示预测成功的概率越高。图 4(d)、(e)、(f) 给出了 3 个符号网络在不同抽样比率下的精确性实验结果。

3.2.2　实验参数设置

给定部分观察的符号网络，符号推断的目标就是通过符号网中已知边符号推断出未知边的符号。本文构建的符号网络模型为有向网络，需要说明的是，所提算法也适用于无向符号网络。实验采用随机抽样的方法将数据集分为训练集 (training data set) 和测试集 (testing data set)。训练集被看作部分观测的符号网络，利用测试集训练模型，然后对测试集中边符号进行预测；测试集分别为整个符号网络的 15%, 30%, …, 90%。对于 MF 和 MF-Bias 算法，首先需要对矩阵$\boldsymbol{P}$、$\boldsymbol{Q}$进行初始化，这里我们令其为全 1 矩阵，有时也将$\boldsymbol{P}$、$\boldsymbol{Q}$的初始值设为随机矩阵。另外，模型还需要确定 3 个参数，即κ、α和λ，其中κ为符号网络的秩，取$\kappa=5$；λ为惩罚因子，取$\lambda=0.12$；α是学习速度，初始值取$\alpha=0.2$，且每次迭代后使α值衰减 (α*=0.9)，目的是使算法尽快收敛，最大迭代次数为 30 次。得到的预测结果是符号网上两节点间以正或负的符号相连的倾向，这一预测值并不是离散的±1 而是连续的值，因此得到预测结果后需要对预测结果进行划分，划分方法有直接划分、全局划分、局部划分、从众划分[14]，本文采用直接划分，即预测结果≥0 则预测符号为正，否则为负。以下通过均方根误差 (RMSE) 和预测精确性 (Accuracy) 评价各算法的预测效果。

3.2.3　实验结果及讨论

图 4 的实验结果显示：随着抽样数据的增加，预测误差 (RMSE) 减小，预测精确度增加。基于低秩矩阵分解的方法 (包括 MF、MF-Bias) 获得了比其他算法更好的预测效果，这说明在符号网络中节点标注的偏置现象确实存在，同时，由于 MF-Bias 充分考虑了节点的局部偏置特性，得到了相较于基本矩阵分解算法好的预测精度，例如：在数据集 Epinions 上当训练集为 90% 时，预测精确度为 95.04%，相较于基本矩阵分解方法提高了 0.6%，LR 方法提高了 2.3%，在其他两个数据集上也得到了与图 4(a) 相似的结果 (见图 4(b)~(c))，RMSE 误差分析结果 (见图 4(d)~(f)) 与预测精确度得到相似的结论：本文所提 MF-Bias 模型获得了最小预测误差。实验表明：带有偏置的矩阵分解方法能够很好地对抗数据稀疏带来的问题并提高预测效果。尽管两种启发式算法 (ID 和 OD) 的预测精度都低于矩阵分解模型和逻辑回归模型，但是它们的特点是计算复杂度低，因为它们仅仅使用待预测边两端节点的局部信息且能在一定程度上反映数据的结构特性。不同数据集 ID 和 OD 的效果截然不同，在 Slashdot 上 OD 好于 ID，在 Wiki 上 ID 好于 OD，可见仅考虑出度或入度作为预测依据不够合理，用户在不同数据集上的行为特征值得进一步思考。

3.3　秩与预测精度

根据 1.3 节可知，符号网络邻接矩阵的秩与弱平衡结构间存在必然联系：当符号网络满足κ-平衡条件时，网络节点可以被分成κ个子集，邻接矩阵具有低秩性且矩阵的秩恰好等于κ，而矩阵分解的本质是做降维操作，将会把邻接矩阵分解为 2 个κ行n列的矩阵，那么到底将邻接矩阵分解为多少行合适呢？本文分别令κ=1、2、4、5、6、7、8、16、32，对两种矩阵分解算法在 3 个数据集进行精确度测试，所有实验均取测试集为 90%，其余各参数与 3.2 节相同。实验结果如图 5 所示，实验结果显示：相较于κ=1，κ=2 时预测精度有大幅提高，这支持了 1.4 节所述结构平衡理论的正确性，κ值从 2~5 预测精度

有较大幅度提高, 大部分数据集在κ=7 时预测精度达到最优 (Slashdot 数据集在κ=5 时精确度最优), $\kappa\geq7$ 后预测精度变化不大。实验结果与 Chiang 等在文献[13]中得到的结论一致, 也证明符号网络邻接矩阵的低秩特性明显。实验还发现: 比起基本矩阵分解算法, 带偏置的矩阵分解算法对κ值更加鲁棒, 即随着κ的变化符号预测的精确度变化不大, 这是因为在带偏置的矩阵分解模型中节点及其邻居的特征与低秩特性共同决定模型的精确度, 因此获得了较高的精确度, 这也证明节点的局部特性对预测效果有影响。

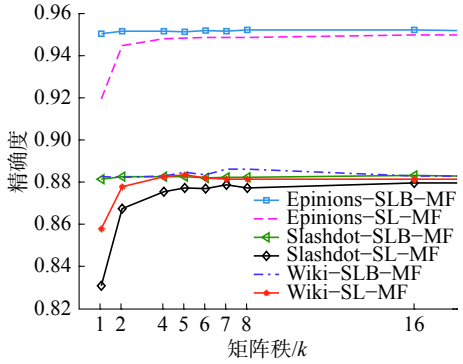


图 5　预测精度与矩阵的秩κ间的关系

Fig. 5　The relation between prediction accuracy and κ

4　结束语

真实的复杂系统中对立关系普遍存在, 利用符号网络对这些复杂系统建模能够很好地表达节点间的对立关系, 符号属性对分析、理解复杂网络的拓扑结构、功能、动力学行为具有十分重要的理论意义。要利用符号属性, 首要的问题就是对未知边符号的正确预测。本文对已有的符号网络预测方法进行了分类和总结。为了同时利用节点的局部特征和全局特征进行符号预测, 在基于利用网络全局特征的低秩矩阵分解方法的基础上改写优化目标函数使之能够描述待预测边两端节点的出度和入度局部特征, 给出了带有偏置的低秩矩阵分解方法。实验结果证实: 添加节点局部特征后的低秩矩阵分解方法能够得到较其他基准算法好的预测效果, 且互惠信息能够进一步提高预测精度。

未来符号预测的研究方向会向两个不同方向发展: 1) 进一步利用丰富的元数据信息, 因为元数据蕴含了用户间的熟识度、声誉、语义与态度等重要信息, 元数据可以在缺少结构信息时保证预测精度, 当然付出的代价是模型复杂度升高, 运算速度降低; 2) 降低模型复杂度以适应于数量巨大的在线符号网络挖掘, 此时基于网络局部信息的符号预测方法具有优势, 因为这类算法易于被并行化, 从而提高运算速度, 当然负面影响会带来一定预测效果的下降。如何充分利用局部信息的研究还显得不足, 如节点间除了出度和入度还有哪些连接特点能用结构平衡理论或地位理论来解释。当节点间的嵌入性很低时结构平衡等社会学理论将失效, 怎样保证预测的精确度?

另外, 还需进一步丰富符号网络结构的理论研究, 目前用于符号预测的理论只有结构平衡理论和地位理论, 近年并未有较大突破。也就是说, 对符号网络结构演化、动力学行为的分析仍然不能解释符号网络结构的形成, 从而制约了符号预测方法的进一步发展, 根据 3.3 节的研究发现本文所提算法对矩阵的秩鲁棒, 及在秩取 5 和 7 时预测效果最好, 这一结果的深层社会学理论含义及其与符号网络形成机制间的联系将另文讨论。这也是下一步将要研究的内容。

符号网络作为近年来从基本复杂网络衍生出的新网络模型和符号预测问题作为新模型上的新问题, 人们对它们的理解还远远不够, 符号网络的拓扑结构、动力学行为以及在个性化推荐、态度预测、用户特征分析与聚类等方面的应用将会受到更多的研究和关注。

参考文献:

[1] 程苏琦, 沈华伟, 张国清,等. 符号网络研究综述[J]. 软件学报, 2014, 25(1): 1–15.
CHENG Suqi, SHEN Huawei, ZHANG Guoqing et al. Survey of signed network research[J]. Journal of software, 2014, 25(1): 1–15.

[2] TANG Jiliang, AGGARWAL C, LIU Huan. Recommendations in signed social networks[C]//Proceedings of the 25th International Conference on World Wide Web. Montréal, Canada, 2016: 31–40.

[3] LI Dong, XU Zhiming, CHAKRABORTY N, et al. Polarity related influence maximization in signed social networks[J]. PLoS one, 2014, 9(7): e102199.

[4] EVERETT M G, BORGATTI S P. Networks containing negative ties[J]. Social networks, 2014, 38: 111–120.

[5] HEIDER F. Attitudes and cognitive organization[J]. The Journal of psychology: interdisciplinary and applied, 1946, 21(1): 107–112.

[6] SZELL M, LAMBIOTTE R, THURNER S. Multirelational organization of large-scale social networks in an online world[J]. Proceedings of the national academy of sciences of the United States of America, 2010, 107(31): 13636–13641.

[7] CHU Lingyang, WANG Zhefeng, PEI Jian, et al. Finding gangs in war from signed networks[C]//Proceedings of the

22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 1505–1514.

[8] DAVIS J A. Clustering and structural balance in graphs[J]. Human relations, 1967, 20(2): 181–187.

[9] LESKOVEC J, HUTTENLOCHER D, KLEINBERG J. Signed networks in social media[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Atlanta, USA, 2010: 1361–1370.

[10] GUHA R, KUMAR R, RAGHAVAN P, et al. Propagation of trust and distrust[C]//Proceedings of the 13th International Conference on World Wide Web. New York, USA, 2004: 403–412.

[11] LESKOVEC J, HUTTENLOCHER D, KLEINBERG J. Predicting positive and negative links in online social networks[C]//Proceedings of the 19th International Conference on World Wide Web. Raleigh, USA, 2010: 641–650.

[12] WANG Guannan, GAO Hui, CHEN Lian, et al. Predicting positive and negative relationships in large social networks[J]. PLoS one, 2015, 10(6): e0129530.

[13] CHIANG K Y, NATARAJAN N, TEWARI A, et al. Exploiting longer cycles for link prediction in signed networks[C]//Proceedings of the 20th ACM international Conference on Information and Knowledge Management. Glasgow, Scotland, UK, 2011: 1157–1162.

[14] 蓝梦微, 李翠平, 王绍卿, 等. 符号社会网络中正负关系预测算法研究综述[J]. 计算机研究与发展, 2015, 52(2): 410–422.
LAN Mengwei, LI Cuiping, Wang Shaoqing, et al. Survey of sign prediction algorithms in signed social networks[J]. Journal of computer research and development, 2015, 52(2): 410–422.

[15] SONG Dongjin, MEYER D A. Link sign prediction and ranking in signed directed social networks[J]. Social network analysis and mining, 2015, 5: 52.

[16] KUNEGIS J, SCHMIDT S, LOMMATZSCH A, et al. Spectral analysis of signed graphs for clustering, prediction and visualization[C]//Proceedings of the 2010 SIAM International Conference on Data Mining. Columbus, USA, 2010: 559–570.

[17] HSIEH C J, CHIANG K Y, DHILLON I S. Low rank modeling of signed networks[C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 507–515.

[18] MENON A K, ELKAN C. Link prediction via matrix factorization[C]//Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases. Athens, Greece, 2011: 437–452.

[19] AGRAWAL P, GARG V K, NARAYANAM R. Link label prediction in signed social networks[C]//Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 2591–2597.

[20] CARTWRIGHT D, HARARY F. Structural balance: a generalization of Heider's theory[J]. Psychological review, 1956, 63(5): 277–293.

[21] CANDÈS E J, TAO T. The power of convex relaxation: near-optimal matrix completion[J]. IEEE transactions on information theory, 2010, 56(5): 2053–2080.

作者简介：

苏晓萍，女，1971 年生，教授，主要研究方向为复杂网络上动态信息传播、信息检索。主持省部级项目 2 项，发表学术论文 10 余篇。

宋玉蓉，女，1971 年生，教授，博士生导师，博士，主要研究方向为复杂网络上动态信息传播、网络控制与优化。主持国家自然科学基金项目 2 项、教育部人文社科项目 1 项，发表学术论文 30 余篇，被 SCI、EI 收录多篇。