基于 PageRank 的社交网络影响最大化传播模型与算法研究

宫秀文[1] 张佩云[1,2]
(安徽师范大学数学计算机科学学院 芜湖 214003)[1]
(中国科学技术大学计算机科学与技术学院 合肥 230026)[2]

摘 要 社交网络中影响最大化问题是指找出最具有影响力的 k 个节点,使得最终社交网络中被影响的节点最多,信息传播范围最大。针对影响最大化问题,目前已存在一些基本传播模型,但是这些模型没有考虑网络中节点的相关性和重要性,而网络中节点的相关性和重要性是衡量其影响力的一个重要指标,因此,提出了一种基于网页排名算法的信息传播模型(PageRank-based Propagation Model,PRP),然后利用贪心算法来近似求解影响最大化问题。实验结果表明,基于 PageRank 的传播模型解决影响最大化问题的效果比传统的线性阈值模型、加权级联模型和独立级联模型的效果更好,影响力范围更大。

关键词 社交网络,影响最大化,PageRank,信息传播模型与算法
中图法分类号 TP311 **文献标识码** A

Research on Propagation Model and Algorithm for Influence Maximization in Social Network Based on PageRank

GONG Xiu-wen[1] ZHANG Pei-yun[1,2]
(School of Mathematics and Computer Science, Anhui Normal University, Wuhu 214003, China)[1]
(School of Computer Science, University of Science and Technology of China, Hefei 230026, China)[2]

Abstract The influence maximization problem in social network is to find top-k influential nodes in graph that maximize the number of influenced nodes. Some basic propagation models have been proposed to solve the influence maximization problem. But those models do not consider the relativity and importance of the node which we consider as an important measurement of influence. Thus, we propose a new PageRank-based propagation model, and employ the Greedy Algorithm to solve the influence maximization problem. The experimental results show that our proposed model is more effective than traditional Linear Threshold Model, Weighted Cascade Model and Independent Cascade Model in solving the influence maximization problem.

Keywords Social network, Influence maximization, PageRank, Information propagation models and algorithm

1 引言

社交网络影响力[1]是指人们接受他人信息传播的过程。在该领域中,Domingos 和 Richardson 提出了社交网络影响最大化问题[2],用图来表示社交网络。我们的目标是在图中找出最具有影响力的 k 个节点,使得最终社交网络中被影响的节点最多,信息传播范围最大。信息在社交网络传播过程中都遵循一定的规则,我们称之为信息传播模型。随着社交网络的出现及流行,社交网络影响力成为目前研究的热点。在不同的传播模型基础上,研究影响最大化问题很有意义。目前已存在一些基本传播模型,如线性阈值模型[3](Linear Threshold Model,简称 LT 模型)、独立级联模型[4](Independent Cascade Model,简称 IC 模型)和加权级联模型[5](Weighted Cascade Model,简称 WC 模型)。此外,还有一些重要传播模型,比如热传播模型[6],但是这些模型都没有考虑网络中节点的相关性和重要性,我们认为网络中节点的相关性和重要性是衡量其影响力的一个重要指标,基于此,本文提出了一种基于网页排名算法的信息传播模型(PRP),并在该模型下利用贪心算法来近似求解影响最大化问题。实验表明,在本文提出的模型下解决影响最大化问题的效果比传统的传播模型更好,影响力范围更大。

本文第 2 节介绍相关工作;第 3 节介绍一些理论基础知识;第 4 节介绍提出的 PRP 模型与寻找 Top-k 节点的贪心算法;第 5 节介绍在 4 个不同数据集上进行的实验及结果分析;最后进行工作总结并展望下一步工作。

2 相关工作

社交网络影响力模型问题最近成为社交网络分析的热

本文受国家自然科学基金项目(61201252),安徽省自然科学基金项目(1308085MF100),安徽省高校省级自然科学研究重点项目(KJ2011A128),安徽省科技厅软科学计划项目(11020503009)资助。
宫秀文(1990-),女,硕士生,主要研究方向为数据挖掘与服务计算,E-mail:xwgong.first@163.com;张佩云(1974-),博士后,副教授,硕士生导师,主要研究方向为服务计算、智能信息处理与数据挖掘等。

点，社交网络中一个重要的问题是影响力最大化问题。为了解决影响最大化问题，目前已有一些基本的信息传播模型，下面简要介绍 LT 模型、IC 模型及 WC 模型。

2.1 线性阈值模型

线性阈值模型来源于数学研究，是以接受者为中心的模型。在 LT 模型中，节点 v 的所有活跃父节点 u 以权重 $\omega(u,v)$ 影响子节点 v，且 $\sum_{u\in\Gamma_t(v)}\omega_{u,v}\leqslant 1$，其中 $\Gamma_t(v)$ 表示 t 时刻节点 v 活跃父节点的集合。给定活跃节点的初始集合 A，则信息按照如下过程进行传播：

①对任意节点 $v\in V$，从[0,1]区间随机选择一个阈值 θ_v；

②在传播 t 时刻，所有活跃父节点 u 以权重 $\omega(u,v)$ 影响所有非活跃子节点 v；

③如果 v 的所有活跃父节点对其影响的权重之和大于等于 v 的阈值 θ_v，即，$\sum_{u\in\Gamma_t(v)}\omega_{u,v}\geqslant\theta_v$，那么非活跃节点 v 将在第 $t+1$ 时刻变成活跃节点；

④如果没有更多的节点被激活，那么该传播过程终止。

阈值 θ_v 表示当父节点 u 为活跃节点（该节点接受某个观点或购买了某个商品）时，其子节点 v 同样成为活跃节点的潜在倾向的不同。LT 模型是一个与 0-1 分布有关的概率模型，节点的阈值选取是随机的。对于一个活跃节点初始集合 A（$\subseteq V$），用 $\varphi(A)$ 表示随机激活过程结束时活跃节点的个数，$\varphi(A)$ 是一个随机变量，用 $\delta(A)$ 表示 $\varphi(A)$ 的期望值，我们称 $\delta(A)$ 为初始集合 A 的影响度。

2.2 独立级联模型

独立级联模型是以发送者为中心的模型，是基于概率理论里面的交互粒子系（Particle Systems）设计的一个信息扩散模型。在独立级联模型中，首先为每条有向边 (u,v) 选择一个实数值 $p_{u,v}\in[0,1]$，$p_{u,v}$ 表示 u 通过边 (u,v) 成功影响 v 的概率。给定活跃节点的初始集 A，活跃节点传播过程按照如下规则进行：

①当在第 t 时刻，节点 u 为活跃节点，它只有唯一一次机会激活它的每个非活跃的节点 v，且激活成功的概率为 $p_{u,v}$，如果 u 成功激活 v，则 v 将在 $t+1$ 时刻成为活跃节点；

②如果节点 v 在 t 时刻有多个活跃的父节点 u，则活跃父节点 u 均在 t 时刻以任意顺序尝试激活 v；

③无论 u 是否能够激活 v，在后面的回合中 u 都不能再尝试激活 v；

④信息在整个社交网络中的传播过程一直持续到没有新的激活可能发生为止。

在 IC 模型中，信息通过有向边 (u,v) 传播成功的概率 $p_{u,v}$ 是随机的，对于活跃节点初始集合 $A(\subseteq V)$，用随机变量 $\varphi(A)$ 表示激活过程结束时活跃节点的个数，用 $\delta(A)$ 表示 $\varphi(A)$ 的期望值，我们称 $\delta(A)$ 为初始集合 A 的影响度。

2.3 加权级联模型

在独立级联模型中，激活概率 $p_{u,v}$ 没有考虑节点的度。然而，度较高的节点影响与被影响的概率都较高，基于此，Kempe D，Kleinberg J，Tardos E 提出了加权级联模型，度数高的节点关联的边被赋予较低的激活概率，节点 u 激活节点 v 的概率为 $p_{u,v}=1/d_v$，d_v 表示节点 v 的度。

在加权级联模型中，信息通过有向边 (u,v) 传播成功的概率 $p_{u,v}$ 与节点 v 的度有关，对于活跃节点初始集合 $A(\subseteq V)$，用随机变量 $\varphi(A)$ 表示激活过程结束时活跃节点的个数，用 $\delta(A)$ 表示 $\varphi(A)$ 的期望值，那么我们称 $\delta(A)$ 为初始集合 A 的影响度。

3 理论基础

3.1 影响力最大化问题描述及定义

社交网络（Social Networks）影响最大化问题，是指找到最具有影响力的 k 个节点集合，这些节点能最大范围地将信息传播到社交网络的其他部分。

用有向图 G 表示社交网络，假设在 G 中，初始活跃节点集合为 $A(A\subseteq V)$，集合 A 之外的所有用户节点都是非活跃的，$RS(A)$ 表示社交网络中最终活跃节点的集合，则初始节点集合 A 的影响力范围可以定义如下：$\varphi(A)=|RS(A)|$，$\varphi(A)$ 表示最终活跃的用户节点数目。本文将影响最大化问题表示为一个离散最优化问题，定义如下：在社交网络 G 中，给定参数 k，信息按照特定的传播模型在 G 中传播。找到一个包含有 k 个用户的初始集合 A，使得 A 最终影响的范围最大，即社交网络中最终活跃的用户节点数目最多，亦即 $\varphi(A)$ 最大。

3.2 PageRank 算法

PageRank[7] 称为网页排名，又称网页级别、Google 左侧排名或佩奇排名，是一种由搜索引擎根据网页之间相互的超链接计算的技术，该技术由 Google 创始人拉里·佩奇和谢尔盖·布林于 1998 年在斯坦福大学研发。Google 用它来体现网页的相关性和重要性，而本文正利用网页排名的这个特性来体现社交网络中节点的相关性和重要性。

PageRank 通过网络浩瀚的超链接关系来确定一个页面的等级，把从 A 页面到 B 页面的链接解释为 A 页面给 B 页面投票，根据投票来源（甚至来源的来源，即链接到 A 页面的页面）和投票目标的等级来决定新的等级。简单地说，一个高等级的页面可以使其他低等级页面的等级提升。本文用 PageRank 值来表示节点的相关性和重要性。下面介绍 PageRank 算法，假设一个由 4 个页面组成的小团体：A，B，C 和 D，如果所有页面都链向 A，那么 A 的 PR（PageRank）值就是 B，C，D 的 PageRank 值之和，即 $PR(A)=PR(B)+PR(C)+PR(D)$。假设 B 也有到 C 的链接，并且 D 也有链接到包括 A 的 3 个页面，一个页面不能投票 2 次，所以 B 给每个页面半票。以同样的逻辑，D 投出的票只有三分之一算到了 A 的 PageRank 值上，此时 A 的 PageRank 值为 $PR(A)=\frac{PR(B)}{2}+\frac{PR(C)}{1}+\frac{PR(D)}{3}$，即根据链出页面总数平分一个页面的 PR 值，$L(X)$ 表示从 X 链出页面的数量，$PR(A)=\frac{PR(B)}{L(B)}+\frac{PR(C)}{L(C)}+\frac{PR(D)}{L(D)}$。

为了对那些有链出的页面公平，规定 $q=0.85$（这里的 q 被称为阻尼系数（damping factor），其意义是，在任意时刻，用户到达某页面后并继续向后浏览的概率。$1-q=0.15$ 就是用户停止点击，随机跳到新 URL 的概率）。$q=0.85$ 的 PageRank 算法被用到了所有页面上，表示页面可能被上网者放入书签的概率。下面的算法，没有链入页面的 PageRank 值会是 0，所以 Google 通过数学系统给了每个页面一个初始 PageRank 值。计算公式如式(1)所示。

$$PageRank(p_i)=\frac{1-q}{N}+q\sum_{p_j}\frac{PageRank(p_j)}{L(p_j)} \tag{1}$$

式中，$p_1, p_2, \cdots, p_N$ 是被研究的页面，网络中存在由页面 p_j 指向 p_i 的链接，$L(p_j)$ 是 p_j 链出页面的数量，N 是所有页面的数量。

所有页面的 PageRank 值是一个特殊矩阵中的特征向量，这个特征向量为 $R=\begin{bmatrix} PageRank(p_1) \\ PageRank(p_2) \\ \vdots \\ PageRank(p_N) \end{bmatrix}$，其表示为式(2)所示。

$$R=\begin{bmatrix} (1-q)/N \\ (1-q)/N \\ \vdots \\ (1-q)/N \end{bmatrix}+ q\begin{bmatrix} \ell(p_1,p_1) & \ell(p_1,p_2) & \cdots & \ell(p_1,p_N) \\ \ell(p_2,p_1) & \ddots & & \\ \vdots & & \ell(p_i,p_j) & \\ \ell(p_N,p_1) & & & \ell(p_N,p_N) \end{bmatrix} \tag{2}$$

式中，$\ell(p_i,p_j)=\frac{1}{L(p_j)}$，如果 p_j 不链向 p_i，且对每个 j 都成立，那么 $\ell(p_i,p_j)$ 等于 0，且 $\sum_{i=1}^{N}\ell(p_i,p_j)=1$。

因此，一个页面的 PageRank 值是由其他页面的 PageRank 值计算得到的，PageRank 算法不断重复计算每个页面的 PageRank 值，如果给每个页面一个随机的 PageRank 值(非 0)，那么经过不断地重复计算，这些页面的 PageRank 值会趋向于正常和稳定。

4 PRP 模型与寻找 Top-k 节点的贪心算法

4.1 一种基于 PageRank 的信息传播模型(PRP 模型)

由于传统传播模型没有考虑节点的相关性和重要性，而节点的权威性则是衡量影响力的重要因素，因此，本文提出了一种基于网页排名的传播模型(PRP 模型)。本文用有向图 G 表示社交网络，用 E_G 表示图 G 中边的集合，用 PageRank 值表示节点的权威性，PageRank 值越高的父节点对其子节点的影响力越大。首先计算出图 G 中每个节点的 PageRank 值，用 p_i 表示节点 v_i 的 PageRank 值，p_i^j 表示节点 v_i 第 j 个父节点的 PageRank 值，用 $parent_i$ 表示 v_i 父亲节点的集合，用 ω_{ji} 表示有向边 (v_j,v_i) 上的权重。

我们定义 ω_{ji} 如下，如果有从 v_j 链向 v_i 的边，那么 ω_{ji} 等于 v_j 的 PageRank 值与所有链向 v_i 父节点的 PageRank 值之和的比值；如果没有从 v_j 链向 v_i 的边，那么 ω_{ji} 等于 0。我们可以用式(3)表示 ω_{ji}：

$$\omega_{ji}=\begin{cases} \dfrac{p_i^j}{\sum_{k\in parent_i} p_k}, & (v_j,v_i)\in E_G \\ 0 & (v_j,v_i)\notin E_G \end{cases} \tag{3}$$

在初始时刻，将图中每个节点都置为 0 状态。当 t 时刻，信息开始在社交网络中传播，此时将信息传播源节点的状态都设置为 1，传播源节点将在 $t+1$ 时刻把信息传给其子节点，如果子节点被影响成功，则其状态由 0 变为 1 且不可逆，否则子节点的状态仍为 0。社交网络中有两类节点不会被影响：一类是没有父节点的节点；另一类是从传播源节点开始没有传播路径的节点。而在社交网络中，有 3 类节点是稳定的节点：第一类是传播源节点，状态始终为 1；第二类是永远不会被影响到的节点，状态始终为 0；第三类是已经被信息影响过的节点，已知该节点是否被成功影响，其状态已确定。如果两个节点之间有环，需要先确定一个节点的状态，另一个节点就可以等其所有父节点全部稳定后确定是否接受信息，从而避免死锁发生。

当在 t 时刻，v_i 所有父节点状态都是稳定的，用 $parent_{i_t}^1$ 表示 t 时刻状态为 1 的所有 v_i 父节点集合，用 $parent_{i_t}^0$ 表示 t 时刻状态为 0 的所有 v_i 父节点集合。那么，在 $t+1$ 时刻，v_i 被其父节点影响成功的概率为 $\sum_{v_j\in parent_{i_t}^1}\omega_{ji}$，即在 $t+1$ 时刻 v_i 状态变为 1 的概率；否则，v_i 没有被影响成功，状态为 0 的概率为 $\sum_{v_j\in parent_{i_t}^0}\omega_{ji}$。因此，$t+1$ 时刻 v_i 的状态可表示成式(4)。

$$f_{t+1}(v_i)=\begin{cases} 1, & p_1=\sum_{v_j\in parent_{i_t}^1}\omega_{ji} \\ 0, & p_0=\sum_{v_j\in parent_{i_t}^0}\omega_{ji} \end{cases} \tag{4}$$

式中，$f_{t+1}(v_i)$ 表示 $t+1$ 时刻 v_i 的状态，p_1 表示 $t+1$ 时刻 v_i 被影响成功的概率，p_0 表示 $t+1$ 时刻 v_i 没有被影响成功的概率。

4.2 寻找 Top-k 节点的贪心算法

贪心算法[8]是由 D. Kempel 提出的，我们基于贪心算法寻找 Top-k 节点来解决影响最大化问题。为了找到种子集 S，一个有效的方法是每一步根据贪心算法的标准确定初始集合中的一个节点，直到找到 k 个节点为止。从空的初始集合开始，每次将使得影响范围函数获得最大边际效益的节点加入初始集合，如算法 1 所示。

算法 1 寻找 Top-k 节点的贪心算法

输入：有向图 G，最终扩散集合大小 k

输出：集合大小为 k 的种子集 S

```
1. 初始化：S=Ø，R=10000
2. for i=1 to k do /* 寻找 k 个种子节点 */
3.    for each vertex v∈V\S do /* 选择边际效益均值最大的节点 */
4.       s_v=0    /* 将边际效益值初始化 */
5.       for t=1 to R do /* 计算 t 时刻节点的边际效益 */
6.          s_v += |RS(S∪{v})| - |RS(S)| /* 结点 v 在所有时刻的
                边际效益总和 */
7.       end for
8.       s_v = s_v / R /* 结点 v 的平均边际效益 */
9.       end for
10.      S = S∪{arg max_{v∈V\S}{s_v}} /* 将边际效益均值最大的节点
             并入初始集合中 */
11. end for
```

算法 1 中，首先定义 $S=\emptyset$；$RS(S)$ 表示集合 S 扩散后社交网络被激活的节点的集合；并定义 $s_v=|RS(S\cup\{v\})|-|RS(S)|$ 表示节点 v 的边际效益影响范围函数。从初始集合 $S=\emptyset$ 开始，每一步都选择使得当前影响范围函数获得最大边际效益的节点，选择策略如下：根据局部最优策略，对集合 $V\backslash S$ 中所有的节点，依次计算 $t=1, t=2, \cdots, t=R$ 时刻节点的边际效益，并对这些时刻的边际效益求均值，最后选择使边际效益均值最大的节点 u，即 $u=\arg\max_{v\in V\backslash S}\{s_v\}$，将 u 并入集合 S 中，即 $S=S\cup\{u\}$。经过 k 步，我们就选择了 k 个影响范围最大的节点。

5 实验和评估

5.1 实验数据集

本文实验是在4个真实数据集上进行的，下面介绍本文使用的4个数据集。

数据集1是个物理领域的合作者网络[4]，节点表示研究者，边表示研究者之间的合作关系，本数据集有10748个节点、53000条边。

数据集2来自社群服务平台Flickr[1]，数据集的实体有用户以及他们的关系，包含用户文件和关系文件，我们的实验从本数据集抽取11328个节点、54870条边。

数据集3来源于Meme Tracker[9]，是一个在线新闻网络，节点表示新闻门户或新闻博客，边表示网站之间的影响关系，本数据集有339936个节点、1574596条边。

数据集4是一个社会新闻分享和投票的网站[2]，数据集里包含不同实体和这些实体之间的联系，我们的实验从本数据集抽取了10536个节点、52400条边。

5.2 实验设计

本实验的基准比较模型是第2节相关工作中提到的线性阈值模型、独立级联模型和加权级联模型。我们基于4.2小节给出的贪心算法，在4个真实数据集上进行实验。设定目标集合大小k分别为0,5,10,15,20,25,30，然后观察本文提出的网页排名传播模型以及基准比较模型的影响范围。

5.3 实验结果

数据集1的实验结果见图1，从图1中可以观察出当k值为15时，网页排名模型的影响范围只比线性阈值模型的影响范围略低，而比另外两个基准模型的影响范围高；当k取其它值时，网页排名模型的影响范围比所有基准比较模型的影响范围都高。

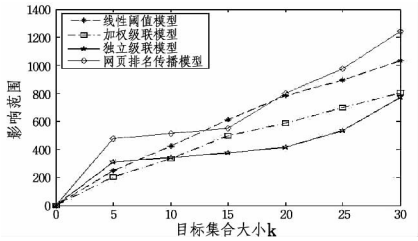

图1 数据集1上不同k影响范围曲线

数据集2的实验结果见图2，从图2中可以观察出当k值等于5时，网页排名模型的影响范围比独立级联模型的影响范围低，比加权级联模型的影响范围高，与线性阈值模型的影响范围不差上下，而当k取其它值时，网页排名模型的影响范围都高于基准比较模型的影响范围。

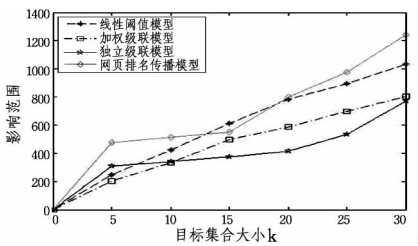

图2 数据集2上不同k影响范围曲线

数据集3的实验结果见图3，从图3中可以观察出当k值等于5时，网页排名模型的影响范围与独立级联模型的影响范围相近，比其它两个基准模型的影响范围高；当k值等于15时，网页排名模型的影响范围与所有基准模型的影响范围都相近；当k取其它值时，网页排名模型的影响范围都高于基准比较模型的影响范围。

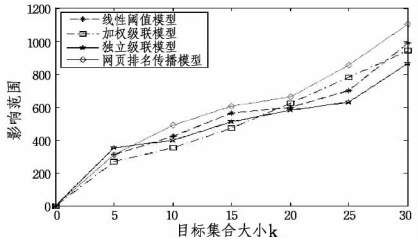

图3 数据集3上不同k影响范围曲线

数据集4的实验结果见图4。从图4中可以观察出当k值等于10时，网页排名模型的影响范围比独立级联模型的影响范围略低，比其它两个基准模型的影响范围都高，而当k取其它值时，网页排名模型的影响范围高于所有基准模型的影响范围。

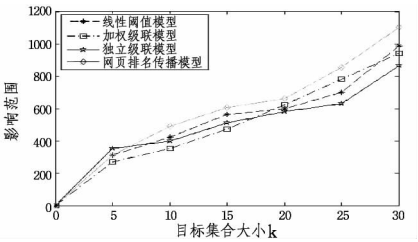

图4 数据集4上不同k影响范围曲线

从以上实验分析得出，网页排名传播模型解决影响最大化问题的效果比传统的线性阈值模型、加权级联模型和独立级联模型的效果更好，影响力范围更大。

结束语 本文基于社交网络中节点的相关性和重要性，针对解决影响最大化问题，提出了一种基于PageRank算法的传播模型(PRP模型)和寻找Top-k节点的贪心算法，以解决影响最大化问题。在4个真实的数据集上进行了实验，结果表明，我们提出的PRP模型与算法比传统的线性阈值、加权级联和独立级联方法的影响力范围更大。尽管本文提出的PRP模型和算法在影响力范围上有所提升，但依然存在值得改进的地方，例如PRP模型只适用于有向网络中信息的传播。我们将进一步研究适用于有向网络以及无向网络的通用信息传播模型。

参 考 文 献

[1] Easley D A, Kleinberg J M. Networks, Crowds, and Markets-Reasoning About a Highly Connected World[M]. Cambridge: Cambridge University Press, 2010

[2] Domingos P, Richardson M. Mining the network value of customers[C]// Seventh International Conference on Knowledge scovery and Data Mining(KDD). 2001:57-66

[3] Mathioudakis M, Bonchi F, Castillo C, et al. Sparsification of Influence Networks[C]// Proceedings of KDD. 2011:529-537

1) http://socialnetworks.mpi-sws.org/data-imc2007.html

2) http://arnetminer.org/heterinf

[4] Kimura M, Saito K, Akano R. Extracting Influential Nodes on a Social Network for Information Diffusion[M]. Data Mining and Knowledge Discovery, 2010: 70-97

[5] Kempe D, Kleinberg J M, Tardos E. Maximizing the spread of influence through a social network[C]// The Ninth International Conference on Knowledge discovery and Data Mining (KDD). 2003: 137-146

[6] Ma H, Yang H, Lyu M R. Mining Social Networks Using Heat Diffusion Processes for Marketing Candidates Selection[C]// Proceedings of CIKM. 2011: 233-242

[7] 金迪，马衍民. PageRank 算法的分析及实现[J]. 计算机应用，2009，18(1001)：118-118

[8] 田家堂，王轶彤，冯小军. 一种新型的社会网络影响最大化算法[J]. 计算机学报，2011，34(10)：1956-1965

[9] Leskovec J, Backstrom L, Kleinberg J M. Meme-tracking and the dynamics of the news cycle[C]// KDD. 2009: 497-506

（上接第 128 页）

表 3 Lymphoma 和 Colon cancer 数据集的约简结果

Data sets	Genes	Samples	GSRS	GSTRS	GSTRSB
Lymphoma	4026	96	7	7	7
Colon cancer	2000	62	6	5	5

实验结果需要比较两方面：一是选择的基因个数；二是分类能力。由表 3 可知，GSRS、GSTRS 和 GSTRSB 算法在 Lymphoma 数据集中提取出的特征基因个数均相同。在 Colon cancer 数据集中，GSTRS 和 GSTRSB 算法比 GSRS 算法提取出的特征基因数少。GSTRS 和 GSTRSB 算法在选取基因个数方面优于 GSRS 算法。

接下来再比较两组基因的分类能力，分别用 KNN，C5.0 作分类器进行实验，并用留一交叉检验，实验结果见表 4。

表 4 Lymphoma 和 Colon cancer 数据集的特征基因分类结果

Data sets	Lymphoma			Colon cancer		
	GSRS	GSTRS	GSTRSB	GSRS	GSTRS	GSTRSB
KNN	93.5%	94.8%	94.8%	79.4%	81.3%	82.9%
C5.0	95.2%	97.4%	97.4%	81.4%	83.6%	85.5%

上述实验结果表明，无论基于粗糙集的基因选择方法还是基于相容关系的基因选择方法提取的基因都能够保持整个基因数据集的分类能力，并且基于相容关系的基因选择方法由于避免了粗糙集离散化过程的信息丢失，提取的特征基因分类精度优于基于粗糙集的基因选择方法提取的基因。根据距离度量函数进一步确定边界域中对象与下近似集均值的接近程度，在 Colon cancer 基因表达数据集中，GSTRSB 方法比 GSTRS 方法能够得到更高的分类准确率。在基因分类能力上 GSTRSB 方法优于 GSTRS 和 GSRS 方法。

结束语 粗糙集不能够处理连续性数据的局限性成为它在基因表达数据研究的主要障碍。本文给出了相容度、综合相容度、相容关系和相容关系下依赖度的定义，提出了基于相容粗糙集的基因特征选择方法，进一步给出了距离度量函数，提出了基于相容粗糙集的改进的基因特征选择方法，并通过一个实例说明了本文提出的方法。由于本文提出的方法避开了离散化过程，并根据距离度量函数确定了边界域中的不确定信息，因此减少了信息损失，从而相对于基于粗糙集理论的基因特征选择方法选择的基因有更好的准确率。本文提出的方法为基因组的基因选择研究提供了一种新的尝试。

参 考 文 献

[1] Tibshirani R, Hastie T, Narashiman B, et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression[C]// Nat'l Academy of Sciences. USA, 2002: 6567-6572

[2] Kohavi R, John G H. Wrappers for feature subset selection[J]. Artificial Intelligence, 1997: 273-324

[3] Banerjee M, Mitra S, Banka H. Evolutinary-rough feature selection in gene expression Data[J]. IEEE Transaction on Systems, Man, and Cybernetic, Part C: Application and Reviews, 2007, 37: 622-632

[4] Momin B F, Mitra S, Datta Gupta R. Reduct generation and classification of gene expression data[C]// Proceeding of First International Conference on Hybrid Information Technology (ICHICT06). New York, 2006: 699-708

[5] Pawlak Z. Rough sets[J]. International Journal of Information Computer Science, 1982, 11(5): 341-356

[6] Dubois D, Prade H. Putting rough sets and fuzzy sets together [J]. Intelligent Decision Support, 1992: 203-232

[7] Jensen R, Shen Q. Tolerance-based and fuzzy-rough feature selection[C]// Proceedings of the 16th International Conference on Fuzzy Systems (FUZZ-IEEE07). 2007: 877-882

[8] Liang J Y, Li R. Distance: A more comprehensible perspective for measures in rough set theory[J]. Knowledge-Based Systems, 2012, 27: 126-136

[9] Parthaláin N M, Shen Q. Exploring the boundary region of tolerance rough sets for feature selection[J]. Pattern Recognition, 2009, 42: 655-667

[10] Yao Y Y, Yao B X. Covering based rough set approximations [J]. Information Sciences, 2012, 200: 91-107

[11] 苗夺谦，胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展，1999，36(6)：681-684

[12] Yang X B, Xie J, Song X N, et al. Credible rules in incomplete decision system based on descriptors[J]. Knowledge-Based Systems, 2009, 22: 8-17

[13] Shen Q, Chouchoulas A. A rough-fuzzy approach for generating classification rules[J]. Pattern Recognition, 2002, 5: 2425-2438

[14] Ou Yang Y P, Shieh H M, et al. Combined rough sets with flow graph and formal concept analysis for business aviation decision-making[J]. Journal of Intelligent Information Systems, 2011, 36 (3): 347-366

[15] 王国胤. Rough 集理论与知识获取[M]. 西安：西安交通大学出版社，2001

[16] 苗夺谦. 粗糙集理论中连续属性的离散化方法[J]. 自动化学报，2001，27(3)：296-302

[17] Grzymala-Busse J W. Discretization of numerical attributes[M]. Klösgen W, Zytkow J, ed. Handbook of Data Mining and Knowledge Discovery, Oxford University Press, 2002: 218-225

[18] Golub T R, Slonim D K, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring[J]. Science, 1999, 286: 531-537

[19] Wang L P, Feng C, Xie X. Accurate cancer classification using expressions of very few genes[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2007, 4: 40-53

[20] Grzymala-Busse J W, Grzymala-Busse W J. Handling missing attribute values[M]. Maimon O, Rokach L, ed. Handbook of Data Mining and Knowledge Discovery, 2005: 37-57