DOI: 10.11992/tis.201806011
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180716.1159.010.html

基于加权聚类集成的标签传播算法

张美琴[1], 白亮[2], 王俊斌[1]
（1. 山西大学 计算机与信息技术学院, 山西 太原 030006; 2. 山西大学 计算智能与中文信息处理教育部重点实验室, 山西 太原 030006）

摘　要: 标签传播算法 (LPA) 是一种高效地处理大规模网络的社区发现算法, 由于其近乎线性的时间复杂度而受到广泛关注。然而, 该算法每个节点的标签依赖于其邻居节点, 其迭代速度和聚类有效性对标签信息的更新顺序非常敏感, 影响了社区发现结果的准确性和稳定性。基于该问题, 提出了一种基于加权聚类集成的标签传播算法。该算法利用多次标签传播算法的结果作为基聚类集, 并用模块度评估每个基聚类的重要性, 使其作为节点相似性度量的权值形成加权相似性矩阵, 最后通过层次聚类得出最终的社区划分结果。在实验分析中, 该算法和其他 5 个具有代表性的标签传播算法的改进算法在真实数据集上进行了比较, 展示了新算法能有效地提高标签传播算法的社区发现精度。

关键词: 数据挖掘; 网络数据; 社区发现; 标签传播算法; 聚类集成; 基聚类; 模块度; 加权度量
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)06−0994−05

中文引用格式: 张美琴, 白亮, 王俊斌. 基于加权聚类集成的标签传播算法[J]. 智能系统学报, 2018, 13(6): 994–998.
英文引用格式: ZHANG Meiqin, BAI Liang, WANG Junbin. Label propagation algorithm based on weighted clustering ensemble[J]. CAAI transactions on intelligent systems, 2018, 13(6): 994–998.

Label propagation algorithm based on weighted clustering ensemble

ZHANG Meiqin[1], BAI Liang[2], WANG Junbin[1]
(1. College of Computer Science and Technology, Shanxi University, Taiyuan 030006, China; 2. Key Laboratory of Symbol Computation and Knowledge Engineering (Shanxi University), Ministry of Education, Taiyuan 030006, China)

Abstract: Label propagation algorithm (LPA) is one of the high-efficiency community detection algorithms for processing large-scale network data. It has attracted much attention because of its nearly linear time complexity with the number of nodes. However, in the algorithm, the label of each node depends on the labels of its neighbor nodes, which makes the iteration speed and clustering performance of the algorithm very sensitive to the order of label information update; this influences the accuracy and stability of the community detection result. To solve this problem, a new LPA is proposed based on weighted clustering ensemble. The new algorithm runs the LPAs many times to obtain several partition results, which can be regarded as a base clustering set. Furthermore, the modularity measure is used to evaluate the importance of each clustering. Based on the evaluation results, a weighted similarity measure is defined between nodes to obtain a weighted similarity matrix of pairwise nodes. Finally, hierarchical clustering on the similarity matrix is used to obtain a final community division result. In the experimental analysis, the new algorithm is compared with several other improved LPAs on five real representative network datasets. The experimental results show that the new algorithm is more effective for improving the community detection accuracy.

Keywords: data mining; network data; community detection; label propagation algorithm; clustering ensemble; base clustering; modularity measure; weighted measure

收稿日期: 2018−06−04.　**网络出版日期:** 2018−07−17.
基金项目: 国家自然科学基金项目(61773247).
通信作者: 张美琴. E-mail: landian.zhang@qq.com.

现实世界中存在各种各样的复杂系统, 网络则是这些系统的普遍存在形式, 如人际关系网,

因特网、大型电力网格等。为了能清晰地发现网络中的信息, 需要挖掘网络的社区结构。社区结构是复杂网络的一个重要特性, 整个网络是由若干个“社区”构成的, 每个社区内部的节点之间的连接相对非常紧密, 而各个社区之间的连接却比较稀疏。利用网络的拓扑结构, 能更准确地发现社区。网络社区结构的发现, 有助于更好地了解社区结构、理解网络的功能和特性、预测复杂网络的行为, 在社会网络、信息推荐、精准营销等领域都有很大的应用价值。

目前提出的代表性社区发现算法有潜在空间算法[1]、谱聚类算法[2-3]、模块最大化算法[4-8]、标签传播算法等, 根据不同的科学需要, 这些算法有不同的社区定义或类定义[9]。其中, 由 Raghavan 等[10]在 2007 年提出的标签传播算法 (LPA) 由于其拥有线性时间复杂度而被广泛应用于处理大规模网络。然而, 该算法中每个节点的标签更新取决于其邻居节点, 更新效果受节点初始输入和标签更新顺序的影响。因此 LPA 算法的最终结果存在不确定性, 影响了社区划分的准确性和稳定性。基于 LPA 结果的不稳定性, 众多学者提出了对 LPA 的改进算法[11−18]。例如, Barber 等[11]在 2009 年提出使用模块度最大化的变体作为约束的 LPA 算法 (LPAm), 该算法将标签传播算法转化为等价优化问题处理; Liu 等[12]在 2010 年针对 LPAm 算法可能会出现社区划分陷入局部极大值的问题, 提出一种改进的标签传播算法 (LPAm+), 其核心是将 LPAm 算法与多步贪婪凝聚算法 (MSG) 融合, 最大限度地减少模块空间, 避免出现局部最大值; Xie 等在 2011 年发表的文献[13]中提出了针对提高社区检测速度和提高社区质量的改进 LPA 算法; He 等[14]在 2014 年使用了 PageRank 来衡量网络中节点的重要性, 提出了基于节点重要性的 LPA 算法; Sun 等[15]在 2015 年提出基于中心的标签传播算法 (CenLP), 通过高密度邻居节点的相似性使节点按特定顺序更新标签; Li 等[16]在 2017 年提出改进的 LPA 算法 Stepping LPA-S 算法, 它通过模块度和目标函数 DN 来度量节点或子网的相似性, 其标签通过相似性进行传播, 最终获得了比 LPA 更准确的结果。

虽然已有多种改进的 LPA 算法被提出, 但依然存在稳定性和精确性不高的问题[17−18]。聚类集成技术正是解决聚类算法结果不稳定和不精确的重要方法之一。目前, 多种聚类集成技术也已得到发展[19−21]。因此, 本文通过将聚类集成技术与标签传播算法融合, 提出了基于加权聚类集成的标签传播算法。该算法通过引入模块度来评估每次 LPA 结果的重要性, 加强了每个基聚类的有效性, 最后采用聚类集成技术实现提高社区发现结果的稳定性和准确性。

1　基于加权聚类集成的标签传播算法

1.1　标签传播算法 (LPA)

标签传播算法 (LPA) 的核心思想是每个节点的标签通过其出现次数最多的邻居节点的标签来决定自身的标签, 通过不断迭代, 节点的标签变化达到稳定后, 将最终具有相同标签的节点划分为一个社区, 完成社区划分。其最大的特点是简单、高效, 缺点是每次迭代结果不稳定, 准确率不高。在文献[10]中给出了该算法的具体描述如下。

1) 为所有节点初始化一个唯一的标签, 每个标签代表一个社区。

2) 产生随机遍历顺序遍历所有节点, 每个节点选择其出现次数最多的邻居节点标签来更新自身的标签, 以此达到标签的传播。

3) 若存在节点的邻居节点标签数出现多个最大值, 则随机选其中一个最大值所对应的标签来更新该节点的标签。重复 2), 直到标签更新稳定, 停止迭代。

4) 将具有相同标签的节点划分为一个社区。

该算法的时间复杂度为 $O(n+m)O(n+m)$, 其中, n 代表为结点分配标签的时间, m 代表每次迭代更新的时间。

1.2　基于加权聚类集成的标签传播算法

在介绍本文提出的新算法之前, 首先给出网络、模块度、基聚类集、节点加权相似性度量等相关概念如下。

定义 1　给定$G=\langle V,E\rangle$是一个无向无权的复杂网络, 其中$\boldsymbol{V}=\{v_1,v_2,\cdots,v_n\}$。若$a_{ij}$代表顶点$v_i$与顶点$v_j$的关系, 令$a_{ij}$表示为

$$a_{ij}=\begin{cases}1, & \langle v_i,v_j\rangle\in E\\ 0, & \langle v_i,v_j\rangle\notin E\end{cases} \tag{1}$$

则称$(\boldsymbol{a}_{ij})_{n\times n}$为$G$的邻接矩阵, 记作$\boldsymbol{A}(G)$。

定义 2[6]　设$G=\langle V,E\rangle$是一个含有k个社区的无向无权网络, 令e_{ij}为$k\times k$维矩阵中的元素, 并令$b_i=\sum_j e_{ij}$, 则模块度记作

$$\boldsymbol{Q}=\sum\nolimits_i(e_{ii}-b_i^2) \tag{2}$$

式中: e_{ij}表示网络中连接第i个社区和第j个社区的节点的边数在所有边中所占的比例, b_i表示与第i个社区中的节点相连的边数在所有边中所占的比例。Q值越大, 表示社区划分结果越稳定, 所对应的社区发现算法效果也更具代表性。

定义 3　设$\boldsymbol{X}=\{\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_T\}$是对图$G$中的节

点集V执行T次标签传播算法得到的结果集，其中，$\boldsymbol{P}_t$代表第t次执行标签传播算法的聚类结果。

第t次运行标签传播算法所产生的单个聚类结果为一个基聚类，将多个标签传播算法的结果融合来代替单个标签传播算法的结果，使用聚类集成技术从结果集$\boldsymbol{X}$中发现最一致的结果。然而，聚类集成的结果会受到单个基聚类质量的影响，为了提高聚类结果的稳定性，因此提出加权相似性度量。

定义 4　设$\boldsymbol{P}_t$为基聚类，$\boldsymbol{X}$为基聚类集矩阵，$\boldsymbol{Q}_t$为衡量基聚类$\boldsymbol{P}_t$有效性的模块度值，则节点i与节点j的加权相似性度量定义为

$$\boldsymbol{S}^w(i,j)=\sum_{i,j=1}^{n}\sum_{t=1}^{T}\boldsymbol{Q}_t\sigma(\boldsymbol{X}_{it},\boldsymbol{X}_{jt}) \tag{3}$$

式中：$\sigma(\boldsymbol{X}_{it},\boldsymbol{X}_{jt})=\begin{cases}1, & \boldsymbol{X}_{it}\neq\boldsymbol{X}_{jt}\\ 0, & \boldsymbol{X}_{it}=\boldsymbol{X}_{jt}\end{cases}\quad(i,j\in n,t\in T)$，$\boldsymbol{X}_{it}$为第$t$次标签传播算法中第$i$个节点所对应的标签。

根据定义 4 可知，用模块度$\boldsymbol{Q}$值来衡量每个基聚类的质量，将不同质量的基聚类分配不同的权值，使得新的相似性度量能更有效地反映节点在基聚类空间下的相似性。

基于以上定义，提出基于加权聚类集成的标签传播算法。用 LPA 算法对复杂网络$\boldsymbol{G}$进行T次聚类，聚类产生的结果形成一个$n\times T$的基聚类集矩阵，然后根据定义 4 融入模块度在基聚类集上求出一个$n\times n$维的节点加权相似性矩阵，最后通过层次聚类算法产生最终的结果。其聚类过程如图 1 所示。

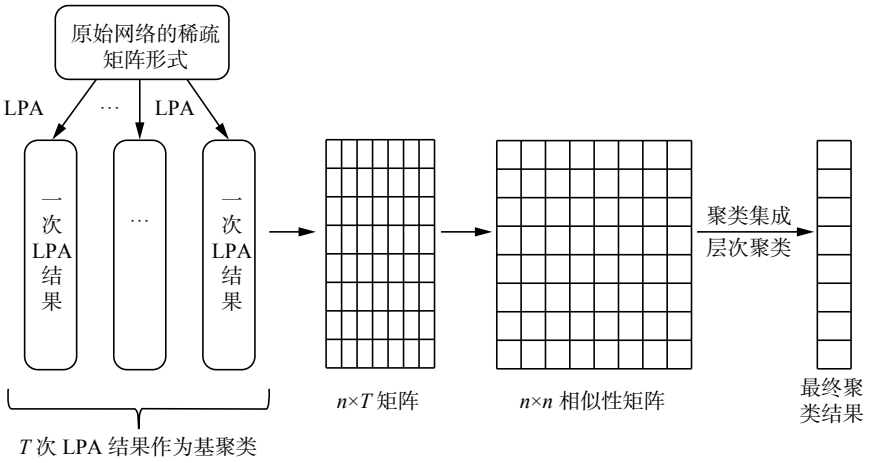

图 1　算法框架

Fig. 1　The framework of the proposed algorithm

具体算法步骤如下：

输入　网络$\boldsymbol{G}$；

输出　节点集$\boldsymbol{V}$的划分结果。

1) 对网络$\boldsymbol{G}$中的节点集$\boldsymbol{V}$执行T次标签传播算法运算，每次产生一列社区划分结果，最终形成$n\times T$的基聚类集矩阵$\boldsymbol{X}$。

2) 根据定义 2 计算每次 LPA 结果的模块度值，$\boldsymbol{Q}_t$来作为对应基聚类的权重。

3) 根据定义 4 计算节点之间的加权相似性矩阵$\boldsymbol{S}^w$。

4) 对加权相似度矩阵$\boldsymbol{S}^w$采用层次聚类方法(linkage) 产生最终的社区划分结果。

2　实验分析

2.1　实验数据

为了验证本文提出的算法的有效性，选取了 5 个不同规模的真实网络数据集，分别为 football、dolphins、karate、web、word 数据集。其中 football 数据集是由美国橄榄球联赛中球队的比赛关系构成的网络，共包含 115 支球队。Dolphins 数据集是由新西兰的一个海豚群体组成的海豚关系网，网络中的边表示两只海豚之间的频繁接触次数。Karate 数据集是一个由 34 个空手道俱乐部成员之间的关系构成的社会网络，网络中的边表示两个俱乐部成员经常出现在俱乐部活动之外的其他场合的次数。Web 数据集是某网站上所有网页构成的关系网，节点代表网页，边代表超链接。Word 数据集是名词形容词搭配网络。数据集的信息描述如表 1 所示。

表 1　数据集描述

Table 1　Description of network data sets

数据集	#vertex	#edges	#Classes
football	115	613	12
karate	34	78	2
dolphins	62	159	2
word	112	425	2
web	75	113	5

2.2　评价指标

本文使用标准互信息 (normalized mutual information, NMI) 和调整兰德系数 (adjusted rand Index, ARI) 来评价最终聚类质量。

标准互信息 (NMI) 和调整兰德系数 (ARI) 常常用于社区划分质量的评价, 能有效衡量给定社区划分结果与真实网络划分的相似程度。NMI 的定义为

$$\mathrm{NMI}(\boldsymbol{A},\boldsymbol{B})=\frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B}c_{ij}\cdot\log\frac{c_{ij}\cdot n}{b_i\cdot b'_j}}{\sum_{i=1}^{c_A}b_i\log(\frac{b_i}{n})+\sum_{j=1}^{c_B}b_{j'}\log(\frac{b'_j}{n})} \tag{4}$$

式中: n代表节点数量, $\boldsymbol{C}$代表混合矩阵, $\boldsymbol{C}$矩阵中的元素$\boldsymbol{C}_{ij}$代表既在真实社区$\boldsymbol{A}$中的i社区又在发现社区$\boldsymbol{B}$中的j社区的节点总数, $c_A(c_B)$是社区$\boldsymbol{A}(\boldsymbol{B})$的社区数目, b_i表示矩阵$\boldsymbol{C}$中第i行元素的总和, $b_{j'}$表示矩阵$\boldsymbol{C}$中第j列元素的总和。对存在真实社区划分的网络, NMI 具有较好的辨识能力, NMI 值越大, 则表明发现的社区结构划分结果越准确, 当发现的社区划分与真实社区完全一致时, NMI 达到最大值 1。

ARI 的定义为

$$\mathrm{ARI}=\frac{r_0-r_s}{\frac{r_1-r_2}{2}-r_s} \tag{5}$$

式中: $r_0=\sum_{ij}\left[\begin{matrix}n_{ij}\\2\end{matrix}\right]$, $r_1=\sum_{i}\left[\begin{matrix}b_i\\2\end{matrix}\right]$, $r_2=\sum_{j}\left[\begin{matrix}b'_j\\2\end{matrix}\right]$, $r_3=\frac{2r_1r_2}{n(n-1)}$。 ARI 的取值范围为[−1,1], 其值越大意味着社区发现结果与真实情况越吻合。

2.3　实验结果分析

为了测试新算法的有效性, 使其分别与 5 个现有的 LPA 的改进算法在真实数据集上进行了比较测试, 这些 LPA 的改进算法包括 LPA[10]、LPA_S[16]、LPAm[11]、LPAm+[12]、BGLL[18]。分别从 NMI 和 ARI 两个评价指标将新算法与经典算法进行了比较, 每个算法在每个数据集上都运行了 100 次, 实验结果如表 2~3 所示。通过对实验结果的对比分析显示, 新算法的实验效果在 football、dolphins、web、word 数据集上都有明显的提高, 即其社区划分更接近于真实社区的划分, 尤其是在 dolphins 数据集, 该算法的效果较其他算法高出 10% 多。虽然在 karate 数据集上新算法的实验结果并非最高, 但也表明新算法在大部分数据集上的表现仍然具有很大优势。同时, 对于算法本身的性能的测评中, 由于该算法只涉及因运行 LPA 算法次数的差异所形成的基聚类集的维度的不同对算法最终结果所造成的影响, 因此本文也对运行不同次标签传播算法得出的最终社区划分结果进行了实验, 实验结果表明, 随着运行次数的增多, 社区划分结果越稳定, 且运行到一定次数时, 社区划分效果均能趋于平稳。综上所述, 本文提出的基于加权聚类集成的标签传播算法较其他算法在 NMI 和 ARI 上都有良好的表现, 且算法本身表现也收敛, 因此新算法能在社区划分的结果上更接近于实际社区划分情况, 提高了社区发现的精确性。

表 2　不同算法的 NMI 值比较

Table 2　Clustering results of different algorithms with respect to NMI

数据集	传播算法	LPA	LPA_S	LPAm	LPAm+	BGLL
Football	0.903	0.704	0.780	0.874	0.893	0.885
dolphins	0.602	0.413	0.533	0.447	0.454	0.445
karate	0.733	0.431	0.837	0.626	0.609	0.587
web	0.152	0.128	0.098	0.135	0.147	0.045
word	0.119	0.085	0	0.021	0.065	0.008

表 3　不同算法的 ARI 值比较

Table 3　Clustering results of different algorithms with respect to ARI

数据集	传播算法	LPA	LPA_S	LPAm	LPAm+	BGLL
Football	0.820	0.425	0.444	0.739	0.811	0.804
dolphins	0.569	0.255	0.318	0.223	0.248	0.280
karate	0.772	0.365	0.882	0.470	0.446	0.462
web	0.016	-0.011	-0.001	0.026	0.037	-0.014
word	0.001	-0.002	0	-0.010	-0.001	-0.009

3　结束语

本文主要利用聚类集成技术对 LPA 进行了改进, 提出了基于加权聚类集成的标签传播算法。该算法首先执行多次 LPA 产生多个标签传播结果作为基聚类集, 并计算出每次 LPA 结果的模块度值作为对应基聚类的权重, 然后计算出融入权值后的节点相似性矩阵, 最后采用层次聚类方法得到最终的社区划分结果。在真实网络数据集上的实验结果表明, 新算法在 NMI 和 ARI 两个评价指标上效果都有所提高, 表明新算法可以获得更好的社区发现结果, 提高了社区发现的精确性。

参考文献:

[1] SARKAR P, MOORE A W. Dynamic social network analysis using latent space models[J]. ACM sigkdd explorations newsletter, 2005, 7(2): 31–40.

[2] ZARE H, SHOOSHTARI P, GUPTA A, et al. Data reduction for spectral clustering to analyze high throughput flow cytometry data[J]. BMC bioinformatics, 2010, 11: 403.
[3] SHI Jianbo, MALIK J. Normalized cuts and image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2000, 22(8): 888–905.
[4] DJIDJEV H N, ONUS M. Scalable and accurate graph clustering and community structure detection[J]. IEEE transactions on parallel and distributed systems, 2013, 24(5): 1022–1029.
[5] DUCH J, ARENAS A. Community detection in complex networks using extremal optimization[J]. Physical review E, 2005, 72(2): 027104.
[6] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physical review E, 2004, 69(6): 066133.
[7] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of the national academy of sciences of the United States of America, 2002, 99(12): 7821–7826.
[8] NEWMAN M E J. Modularity and community structure in networks[J]. Proceedings of the national academy of sciences of the United States of America, 2006, 103(23): 8577–8582.
[9] TANG Lei, WANG Xufei, LIU Huan. Community detection via heterogeneous interaction analysis[J]. Data mining and knowledge discovery, 2012, 25(1): 1–33.
[10] RAGHAVAN U N, ALBERT R, KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical review E, 2007, 76(3): 036106.
[11] BARBER M J, CLARK J W. Detecting network communities by propagating labels under constraints[J]. Physical review E, 2009, 80(2): 026129.
[12] LIU X, MURATA T. Advanced modularity-specialized label propagation algorithm for detecting communities in networks[J]. Physica A: statistical mechanics and its applications, 2010, 389(7): 1493–1500.
[13] XIE Jierui, SZYMANSKI B K. Community detection using a neighborhood strength driven label propagation algorithm[C]//Proceedings of 2011 IEEE Network Science Workshop. West Point, NY, USA: IEEE, 2011: 188–195.
[14] HE Miao, LENG Mingwei, LI Fan, et al. A node importance based label propagation approach for community detection[M]//SUN Fuchun, LI Tianrui, LI Hongbo. Knowledge Engineering and Management: Proceedings of the Seventh International Conference on Intelligent Systems and Knowledge Engineering, Beijing, China, Dec 2012 (ISKE 2012). Berlin Heidelberg: Springer, 2014: 249–257.
[15] SUN Heli, LIU Jiao, HUANG Jianbin, et al. CenLP: a centrality-based label propagation algorithm for community detection in networks[J]. Physica A: statistical mechanics and its applications, 2015, 436: 767–780.
[16] LI Wei, HUANG Ce, WANG Miao, et al. Stepping community detection algorithm based on label propagation and similarity[J]. Physica A: statistical mechanics and its applications, 2017, 472: 145–155.
[17] ŠUBELJ L, BAJEC M. Unfolding communities in large complex networks: combining defensive and offensive label propagation for core extraction[J]. Physical review E, 2011, 83(3): 036103.
[18] BLONDEL V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008(10): 10008.
[19] ELHADARY R S, TOLBA A S, ELSHARKAWY M A, et al. An efficient and robust combined clustering technique for mining in large spatial databases[C]//Proceedings of 2007 International Conference on Computer Engineering and Systems. Cairo, Egypt, 2007: 439–445.
[20] FRED A L N, JAIN A K. Data clustering using evidence accumulation[C]//Proceedings of the 16th International Conference on Pattern Recognition. Quebec City, Quebec, Canada, 2002: 276–280.
[21] YANG Yan, KAMEL M S. An aggregated clustering approach using multi-ant colonies algorithms[J]. Pattern recognition, 2006, 39(7): 1278–1289.

作者简介：

张美琴，女，1992 年生，硕士研究生，主要研究方向为社区检测。

白亮，男，1982 年生，副教授，博士，主要研究方向为数据挖掘、机器学习。

王俊斌，男，1994 年生，硕士研究生，主要研究方向为数据挖掘。