半监督模式下社团结构划分方法

孔 健,谢福鼎,孙 岩
KONG Jian,XIE Fu-ding,SUN Yan

辽宁师范大学 计算机与信息学院,辽宁 大连 116081
College of Computer and Information Technology,Liaoning Normal University,Dalian,Liaoning 116081,China
E-mail:xiefd@sohu.com

KONG Jian,XIE Fu-ding,SUN Yan.Classification algorithm based on semi-supervised learning.Computer Engineering and Applications,2009,45(23):158-161.

Abstract: An information transfer classification algorithm based on semi-supervised learning is proposed in this paper to partition the network with labeled nodes and unlabeled nodes into different clusters.The classification parameters of all unlabeled nodes in network are firstly determined in terms of the suggested approach,and the clustering results can be obtained by iteratively computing these parameters.The analysis of experimental data has proven that this algorithm has better effect on semi-supervised classification.

Key words: information transfer;semi-supervised classification;clustering;complex networks;community structure

摘 要: 为了对有标签和无标签节点混合的网络进行分类,给出了一种基于半监督学习的信息传递分类算法,算法首先确定网络中无标签节点的分类参数,然后通过对网络中所有无标签节点进行有限次的迭代计算,可以对所有节点进行分类。实验数据分析证明了该算法在进行半监督分类时具有比较好的效果。

关键词: 信息传递;半监督分类;聚类;复杂网络;社团结构
DOI: 10.3778/j.issn.1002-8331.2009.23.044 **文章编号:** 1002-8331(2009)23-0158-04 **文献标识码:** A **中图分类号:** TP18

1 引言

随着对网络性质的物理意义和数学特性的深入研究,人们发现许多实际网络都具有一个共同的性质,即社团结构,也就是说,整个网络是由若干“群(group)”或“团(cluster)”构成的。为了找出网络中的社团结构并对其进行深入研究,人们设计了许多的算法来寻找网络中的社团结构,经典的算法如:Kernighan-Lin 算法[1]:他采用了一种贪婪算法,根据社团内部及社团间的边最优化的原则对原始的网络进行分类;基于 Laplace 矩阵特征值的谱平分法[2-3],另外分级聚类也是寻找社会网络中社团结构的一类传统算法,包括凝聚方法和分裂方法[4]。近年来,关于如何寻找复杂网络的理想社团结构仍是计算机科学的研究热点之一。

然而在寻找复杂网络的社团结构时,几乎所有的算法都属于无监督学习的范畴。但有时,在一个网络中有少数节点的类标签会被标记出来,如何利用这些具有类标签的节点找出社团结构,在传统的寻找社团结构的算法中并没有涉及到,半监督学习恰恰就是运用有标签和无标签节点进行分类的一种机制,利用这一机制可以解决如何通过有标签和无标签节点找出网络中的社团结构这个问题。

在传统的监督学习中,学习器通过对大量有标记的(labeled)训练例进行学习,建立模型用于预测未见示例的标记。这里的“标记”(label)是指示例所对应的输出,在分类问题中标记就是示例的类别。而半监督学习是从监督学习的角度出发,考虑带标签训练样本不足时如何利用大量无标签样本信息辅助分类器的训练。针对基于半监督学习的聚类和分类,前人已经研究得到了许多不同的方法,根据对标注数据和未标注数据的不同利用方法来分类,则半监督学习可分为半监督分类和半监督聚类。这些算法利用大量未标注数据辅助监督学习过程,改善分类结果[5]。包括 co-training[6]、transductiveSVMs[7],以生成式模型为分类器的方法,基于图正则化框架的半监督学习算法[8-12]。例如,以生成式模型为分类器的方法,将未标记示例属于每个类别的概率视为一组缺失参数,然后采 EM 算法来进行标记估计和模型参数估计,其代表包括文献[13-15]等。此类算法可以看成是在少量有标记示例周围进行聚类,是早期直接采用聚类假设的做法。其中使用的 EM 算法是一种在有缺失值的情况下计算最大似然估计的迭代算法,它首先对未标注数据进

基金项目: 国家自然科学基金(the National Natural Science Foundation of China under Grant No.10771092);国家重点基础研究发展规划(973)(the National Grand Fundamental Research 973 Program of China under Grant No.2004CB318000);辽宁省教育厅科学技术研究项目(No.2008347);辽宁省科技厅博士启动基金(No.20081079)。

作者简介: 孔健(1983-),男,硕士研究生,主要研究方向为人工智能、知识发现;谢福鼎(1965-),男,博士,教授,通讯作者,主要研究领域为人工智能、数据挖掘;孙岩(1974-),女,博士,副教授,主要研究方向为人工智能、知识发现,贝叶斯理论。

收稿日期: 2009-03-11 **修回日期:** 2009-05-11

行软分类，然后利用获得的完整数据集再进行模型的学习[16]；而 co-training 算法又称为协同训练算法,此类算法隐含地利用了聚类假设或流形假设,它们使用两个或多个学习器,在学习过程中,这些学习器挑选若干个置信度高的未标记示例进行相互标记,从而使得模型得以更新;基于图正则化框架的半监督学习算法，通常先根据训练例及某种相似度度量建立一个图,图中结点对应了(有标记或未标记)示例,边为示例间的相似度,然后,定义所需优化的目标函数并使用决策函数在图上的光滑性作为正则化项来求取最优模型参数，例如,The Graph mincut Learning 算法[17]。

正是由于传统的复杂网络分类算法不能够很好地解决对有标签和无标签节点混合网络的分类问题,因此提出了一种基于半监督学习的分类算法。算法首先确定网络中有标签节点的邻居节点的参数,然后通过对网络中所有无标签节点进行有限次的迭代计算逐步将参数传递到整个网络中的所有节点,并通过分析对所有节点进行分类。

2　算法

该文算法是受 Wu-Huberman 提出的基于电阻网络电压谱的快速谱分割法[18]的启发而提出的,旨在解决当复杂网络中存在有标签节点时的分类问题。假设一个网络 $G=(V,E)$,$|V|=n$,$|E|=m$,G中包含 k 个有标签节点,假设有标签节点可分为 d 类。当有标签节点与其周围节点联系紧密时,它向外传递的信息量就大,反之信息量就会小。在 Wu-Huberman 算法中,这个信息量被具体化为各个节点的电压值。当一个有标签节点向外传递信息时,并不是沿着一条线路传递,而是通过各个线路同时传递,有可能最终会汇聚到某一个无标签节点上,这其实就是一个信息累积的过程。而在半监督学习中,这样的过程则体现得更加明显。而这个用于传递的信息,则可用如下公式计算得到:

$$\theta_i=\frac{|U_i'|}{|U'|}+\frac{|U_i|}{|U|} \tag{1}$$

θ_i表示有标签节点向一个无标签节点传递的关于第 i 类有标签节点的信息量。U_i表示在 G 中与第 i 类节点连接的无标签节点的集合,U_i'则表示在以某一个有标签节点为当前节点以及它的所有邻居节点构成的子网络 G'中,与 i 类节点连接的无标签节点的集合,U 是 G 中与所有 d 类有标签节点相连的无标签节点的集合,U'是 G'中与所有 d 类有标签节点相连的无标签节点的集合。用式(1)对整个网络 G 进行初始化,初始化过程是这样的:计算 G'中无标签节点的 θ 值,即 $\theta_1,\theta_2,\cdots,\theta_i$。每一个参数 θ 由式(1)中两部分的代数和得到,加号前面的部分和加号后面的部分,前者是局部子网络 G'对 θ_i的影响,后者是整个网络 G 对 θ_i的影响。这样做是因为,一个网络中除了局部网络对节点有影响之外,整个网络也对节点存在一定的影响,如果只计算整个网络对 θ_i的影响,那么就不能反映出网络的局部特性,也就是说局部对 θ_i的影响将被淹没。初始有标签节点向其他无标签节点传递的信息则用该类节点的连接数量与所有类连接数量的比值来表示。

用公式(1)进行计算时,(1)如果$|U|=|U_i|=0$时,表示所有节点均为孤立节点,不存在社团结构;(2)如果当确定一个网络 G 之后,式(1)中的后一项将是固定值。假设某一个子网络中,有标签节点个数发生了变动,那么这时在计算这个子网络中的 θ 值时,θ 值就会减小，这时由于当一个子网络中的有标签节点增加时,相应的无标签节点的个数就会减少,这样在 G'中与某一类有标签节点连接的无标签节点的数量就会减少。但这种情况会在这个初始化过程结束后被消除,θ 还是会随着有标签节点的数目增加而增加;(3)当 G 中所有节点都是有标签节点,这时就不存在无标签节点,也就不需要分类。

在初始化过后,每一个和有标签节点直接相连的无标签节点的参数 θ 都被赋予初值。但是这里存在一些同时和若干类有标签节点相连的无标签节点,直观上看,这些节点属于歧义点,它们包含的这些类信息必然会比其他非歧义点要多,从所包含的信息量上可以看出来。当网络 G 初始化完成后开始进行参数 θ 的传递,传递公式如下:

$$\theta_{ji}=\theta_{ji}'+\alpha_{ji} \tag{2}$$

$$\alpha_{ji}=\frac{\theta_{j-1,i}}{h_{j-1}} \tag{3}$$

式(2)中,α_{ji}为第 j-1 个节点向第 j 个节点传递的信息量,θ_{ji}'为第 j 个节点接收 α_{ji}之前的参数值。一个节点包含的信息量是由它原来包含的信息量与从它上一个节点传递过来的信息量的和。它在向下传递时,信息量会有一个衰减的过程,也就是随着传递的进行,传递的信息量会比上个节点包含的信息量少,而α_{ji}的大小则由公式(3)计算得出,其中 h_{j-1}为第 j-1 个节点的度。

整个信息传递的过程伴随着的是每个无标签节点的信息累积的过程,θ_i在增长,但速度不同,随着传递的进行,θ_i中关于各个类的信息量会逐渐分离开来,最终将整个网络分类。一般来说,在一个无标签节点包含的信息中,如果某一类信息量大于其他类信息量时,那么该节点对外表现为信息量大的这一类节点的特征。

例如:存在两个不同类型的网页,当这时存在第三个网页,同时从这两个网页获取信息,并放置在网页上,如果从一个网页获取的信息量要多于从另一个网页获取的信息量,则从直观上来讲该网页必然与前者同属于一个类型。

算法描述:

输入:一个复杂网络 G,分类数为 d,已知标签节点

输出:各个节点的分类结果

步骤 1　从 G 中找出所有有标签节点并从中随机选取 1 个节点作为初始节点。

步骤 2　初始化 G,从步骤 1 中选取的节点开始,选取与其相连接的所有节点及其本身作为一个子图 G',并计算 G'中各数:

$$\theta_i=\frac{|U_i'|}{|U'|}+\frac{|U_i|}{|U|}$$

注意:直到所有有标签节点都被选择过为止,则表示 G 初始化。

结束。

设集合 T 中包含所有无标签节点

(1)$T\neq\phi$ 进入循环;

(2)从 T 中的元素开始逐个计算以当前节点 U_i'及其周围直接连接的节点构成的子图 G'中所有非当前节点的参数值。按照如下规则:

$\theta_{ji}=\theta_{ji}'+\alpha_{ji}$

(3)$T=T-${当前节点}。

步骤 3　计算比较所有无标签节点参数 θ_i并把节点归入对应的 θ 值最大的类中。

3 实验结果及分析

3.1 Zachary 空手道俱乐部关系网络

这个部分针对 Zachary 空手道俱乐部关系网络进行算法的测试,假设已知 1 号节点和 34 号节点为有标签节点,占网络的5.8%,分类数量为 2,如图 1 所示。

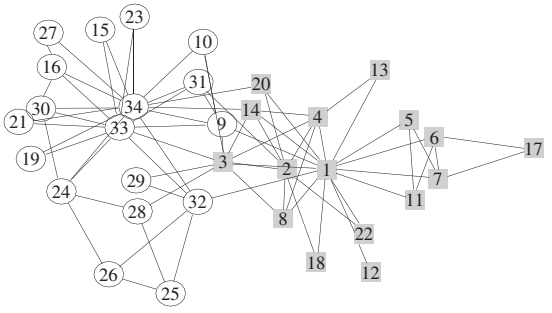

图 1 Zachary 空手道俱乐部关系网络[18]

用该文算法对 Zachary 空手道俱乐部关系网络进行测试,过程如下:首先假设 1 号和 34 号节点为有标签节点,其余节点无标签,然后初始化整个网络,并进行信息的传递,从有标签节点开始逐渐向远处节点传递信息,而每个无标签节点包含的信息也逐渐增加。最终除 3 号节点外,其他节点均正确归类。而当算法采用不同的顺序对网络进行计算时,3 号节点有时同 34 号节点归为一类,有时同 1 号节点归为一类。因为从直观上来看,造成这种现象是由于,3 号节点本身就是一个歧义点,同两个社团连接程度相差无几,因此归为哪一类都有可能。

当算法的初始有标签节点扩大为 3 个时,除 3 号外依然能够正确分类,而当扩大为 4 个有标签节点时,其他节点也可以正确分类。这里可以注意到,当两类有标签节点的已知数量不相等时,算法仍然可以得到正确的分类,也就是说当两类节点的源信息不相等时,仍旧可以正确分类。但是当一类有标签节点的个数与另一类有标签节点的个数只差大于等于 2 时,即当网络中已知 32、33、34 号节点时,20 号节点被误分类。从图 1 中可以看出,20 号节点也属于歧义节点,因此在分类过程中也极有可能被误分类。

下面针对 2 号、3 号、33 号这些具有代表性的节点进行数据分析。之后是所有无标签节点最终的分布图。

图 2~图 4 显示的是第 2 号、3 号、33 号节点参数值随分类过程增长的情况,可以看到,2 号节点包含的参数在算法开始后就开始增长,但是,它包含的参数 θ_2 增长速度远大于 θ_1 的增长速度,随着传递次数的增加,θ_1 与 θ_2 逐渐分离。3 号节点包含的信息量 θ_1 与 θ_2 在算法中前期分离较理想,但是到后期 θ_1 与 θ_2 又有了相等的趋势,原因是 3 号节点本身就是个歧义点,其所包含的关于第一类的信息量和第二类的信息量相差不大,所以就会出现图 3 中的情况。而 33 号节点包含的参数 θ_1 与 θ_2 在算法前期分离度不是很理想,但在中后期,逐渐分离,这种情况出现的原因主要是算法执行次序,或者说是信息传递顺序的问题。物理意义上的解释为,由于 33 号节点属于第一类社团,但在初始阶段信息从第二类社团开始传递,则关于第一类社团的信息并没有传递到 33 号节点,这样就导致了算法初期 θ_1 与 θ_2 的分离度不理想的情况。

图 5 中,显示的是网络 G 中所有无标签节点的最终分布情况。图中的直线为函数 $Y=X$,表示 θ_1 与 θ_2 相等的情况,横纵坐标分别为是 θ_1 和 θ_2。可以非常直观地看出,分属直线两侧的节点各自归为一类,直线附近的节点则为 G 中的歧义点。

表 1 显示的是该文算法与传统的复杂网络分类算法关于时间复杂度和分类准确率的比较。

从表 1 可以看出将这些算法应该用到 Zachary 网络中时,

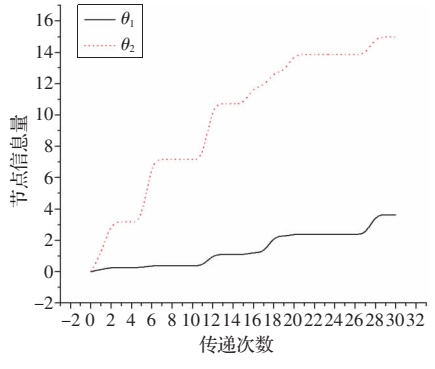

图 2 2 号节点参数变化图

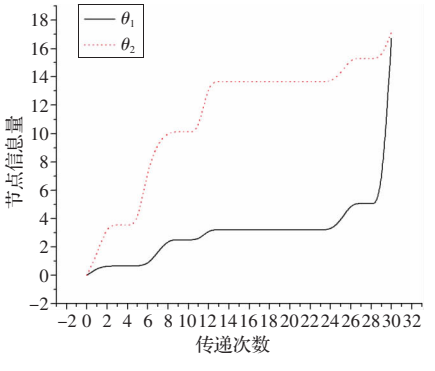

图 3 3 号节点参数变化图

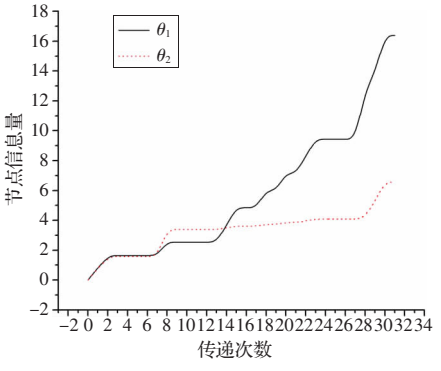

图 4 33 号节点参数变化图

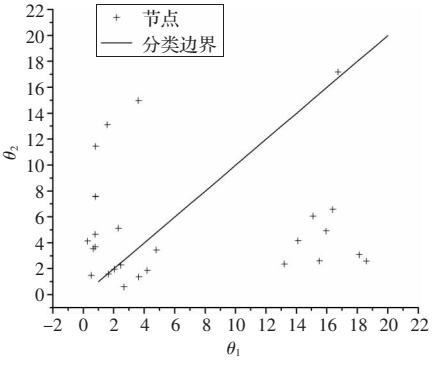

图 5 无标签节点分布图

各算法都表现出了很高的精度,尤其是 Kernighan-Lin 算法,精度是 100%,也就是说它将这个网络中所有节点准确分类,但是它存在一个缺陷,就是必须事先知道两个社团的大小,这使得他在实际网络分析中难以应用,而且它的时间复杂度也比较高。该文的算法在事先知道两个类标签节点的情况下,可以对 Zachary 网络中的所有节点进行比较准确的分类,而且仅为线性时间复杂度 $O(m+n)$。

表 1　算法对比

算法	时间复杂度	准确率/(%)
Wu-Huberman 算法	$O(m+n)$	97
GN 算法	$O(n^3)$	97
Kernighan-Lin 算法	$O(n^2\log n)$	100
该文算法	$O(m+n)$	97
Newman 快速算法	$O(n^2)$	97

3.2　College Football 网络

为了进一步分析该文算法在复杂网络中分类的效果,选取节点数目较多,各节点度分布较均匀的 College Football 网络作为测试对象。它是一个由 115 个节点组成的网络,节点的度为 8~12。已知的分类数目为 12,选取已知类标签的节点个数为 13,占整个网络的 11%。将该文的算法应用到这个网络进行分类,属于 Atlantic Coast 类的 9 个节点、属于 Big Ten 类的 11 个节点、属于 Mountain West 类的 8 个节点、属于 Pacific Ten 类的 10 个节点均正确分类。整个网络中误分类节点一共是 9 个,占整个网络的 7.8%,它们分别是:Baylor、BoiseState、CentralFlorida、LouisianaMonroe、MiddleTennesseeState、Navy、Notre Dame、Rutgers 和 TexasChristian。算法需要的已知类标签节点的个数是随着分类个数变化的,当分类个数不变时,已知类标签节点的个数越多,分类越准确。

表 2 显示的是该文算法与 GN 算法针对 College Football 网络关于时间复杂度和分类准确率的比较。

表 2　同 GN 算法的比较

算法	时间复杂度	准确率/(%)
GN 算法	$O(n^3)$	78
该文算法	$O(m+n)$	92

从表 2 可以看出,该文算法在事先已知极少数目的具有类标签的节点之后,可以通过这极少数的节点与其他无标签节点共同将 College Football 网络分类,分类效果比较好,因此可以看出,在网络中的社团结构不是十分明显的情况下,该文算法依然可以得到较好的分类效果。而 GN 算法比较适合社团结构比较明显的网络的分类,当对 College Football 网络这样社团结构不是十分明显的网络进行测试时,准确率会有所下降。

3.3　Collaboration Network

Collaboration Network[19]是一个 Santa Fe 研究机构的科学家合作网络,选取了其中最大的一个部分进行测试,这个部分由 118个节点组成。假设其中度最大的 4 个节点为已知标签节点,运用该文的算法对这个网络进行分类,最终的准确率为 93%。

从分类结果可以看出,当网络中节点数量比较多,社团结构比较明显时,该文算法可以得到比较好的分类结果。

在对上述三个网络进行测试后,可以看到,当网络有明显社团结构时,该文算法在数据预处理阶段需要的已知类标签节点的个数非常少,在 Zachary 空手道俱乐部关系网络中仅占所有节点的 5.8%,而在 Collaboration Network 中,已知类标签节点的数目仅占所有节点的 3%。当测试 College Football 这样的社团结构不是很明显的网络时,该文算法依然可以得到比较理想的结果。

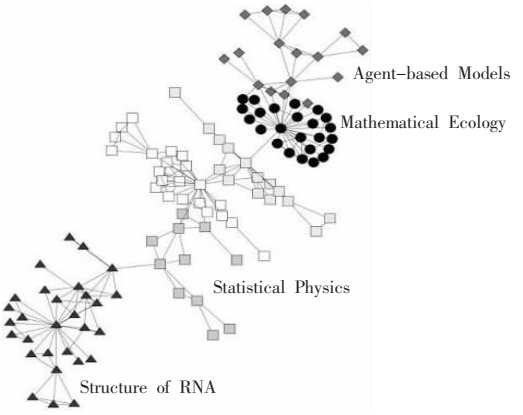

图 6　Collaboration Network

4　结论

该文中的算法利用的是半监督学习模式,借助有标签节点和无标签节点共同得到分类模型,最终将整个网络中的节点分类。而半监督学习的思想是通过数据本身,在未知测试数据的情况下,得到分类模型。它具有一定的扩展性,当有大量新节点加入原始网络时,算法同样可以将这些节点分类,具备半监督学习的"开放"特点。但是该方法存在一个不足之处,即当两类已知节点的数目相差比较大时,算法出现误分情况。为了避免这一情况的发生,需要在数据预处理阶段尽量将各类有标签节点设置为同样数量。这也是今后工作的一个方面,在数据预处理阶段采取一些有效的措施进而避免由于已知节点的数目相差比较大而引起的误分类情况。

参考文献:

[1] Kernighan B W,Lin S.An efficient heuristic procedure for partitioning graphs[J].Bell System Technical Journal,1970,49:291-307.

[2] Fiedler M.Algebraic connectivity of graphs[J].Czech Math J,1973,23.

[3] Pothen A,Simon H,Liou K P.Pattitioning aparse matrices with eigenvectors of graphs[J].SIAM J Matrix Anal Appl,1990,11.

[4] Scott J.Social network analysis:A handbook[M].2nd ed.London:Sage Publications,2002.

[5] Seeger M.Leammg with labeled and unlabeled data[EB/OL].(2002).http://www.dai.ed.ac.uk/~seege/apers.html.

[6] Blum A,Mitchel T.Combining labeled and unlabeled data with co-training[C]//Proceedings of the 11th Annual Conference on Computational Learning Theory,Madison,WI,1998:92-100.

[7] Joachims T.Transductive inference for text classification using support vector machines[C]//Proceedings of the 16th International Conference on Machine Learning,Bled,Slovenia,1999:200-209.

[8] Zhu X,Ghahramani Z,Lafferty J.Semi-supervised learning using Gaussian fields and harmonic functions[C]//Proceedings of the 20th International Conference on Machine Learning(ICML'03),Washington DC,2003:912-919.

[9] Belkin M,Niyogi P.Semi-supervised learning on riemannian manifolds[J].Machine Learning,2004,56(1/3):209-239.

(下转 185 页)