

基于 BP 神经网络的社团分类研究

张兰华^{1,2}, 孙 岩², 薛绍伟², 唐一源²

(1 泰山医学院 信息工程学院, 山东 泰安 271016; 2 大连理工大学 神经信息学研究所, 辽宁 大连 116023)

摘 要: 社团分类对研究复杂网络及其特性具有重要意义, 为此提出将神经网络的分类特性应用于社团分类中, 以 BP 神经网络为模型, 以 Zachary 研究的俱乐部成员关系网络为测试数据, 利用社团节点与神经网络节点之间关系的相似性模拟社团分类结果, 根据成员关系图建立邻接矩阵, 利用邻接矩阵关系进行实验设计建立 BP 神经网络, 通过已知样本学习和检验样本检验, 实现成员关系网络的分类模拟, 数值实验结果表明文章运用的仿真方法是可行的。

关键词: 社团分类; 算法; BP 神经网络; 邻接矩阵; 样本

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2011)06-0197-04

Research of Community Classification Based on Back Propagation Neural Network

ZHANG Lan-hua^{1,2}, SUN Yan², XUE Shao-wei², TANG Yi-yuan²

(1 Department of Information and Engineering, Taishan Medical University, Taian 271016, China;

2 Institute of Neuroinformatics, Dalian University of Technology, Dalian 116023, China)

Abstract: Community classification, an important character of complex network research, can be applied from the advantage of neural network in pattern recognition. The basic idea is that the similarity relationship is introduced to simulate the classification between community and neural network by the model of back propagation neural network and the club membership network that was studied by zachary as the experimental surrounding. The method is to set up the adjacency matrix on the membership, then to build the back propagation neural network on the lab design and the adjacency matrix. The experimental outcomes are testified feasible by studying the known samples and the test samples.

Key words: community classification; algorithm; back propagation neural network; adjacency matrix; sample

1 引言

自然界中许多复杂系统可以通过网络进行描述和表示, 复杂网络是对网络性质物理意义和数学特性的抽象, 除小世界、无标度等重要特征外, 社团结构是研究复杂网络的另一重要特征^[1]. 社团结构在现实中表现为同一社团中的节点更易具有相同或相似性质, 即社团内部节点之间关系相对密切, 社团之间连接相对稀疏, 因此可以通过研究复杂网络的特

性来揭示现实系统的特征和意义, 社团分类在复杂网络应用中, 对于控制疾病及网络病毒的传播等具有重要意义。

由于社团本身的特点, 其度量标准不一, 各种社团分类算法应用环境和突出重点不同, 在分类准确性、算法确定性及时间复杂度方面有不同的结论, 因此社团分类算法致力于寻找正确性高, 确定性强, 简单操作且时间复杂度低的方法。

本文从一个新的角度利用神经网络方法模拟社

收稿日期: 2010-09-25; **修回日期:** 2010-12-12

基金项目: 国家自然科学基金项目(60971096); 山东省高等学校科技计划项目(J10LC59); 泰山医学院科研计划课题青年基金项目(2009ZRQN053); 泰山医学院科研计划课题面上项目(2010ZR100)

团分类,以 Zachary 研究的空手道俱乐部内部成员关系网络为测试数据,以反向传播(Back Propagation;BP)神经网络为基础,分析社团分类与 BP 网络之间的相似特性,通过对条件加以限制,建立符合 BP 学习算法的社团分类神经网络,利用邻接矩阵的数据关系简单有效的完成社团分类,给出详细求解过程和实验分组设计与仿真结果,并对分类结果进行分析和总结。

2 研究现状

2.1 社团分类

社团通常与模块、类、集群、集团、组等 synonym^[2],在网络中可描述为网络的一个子图,把满足同一主题节点及节点之间的关系组成的网络看作整个网络的子图。子图具有各种现实含义,比如社会关系中具有相同兴趣的社会团体,论文网络中对同一主题引用的相关论文,因特网中对同一主题感兴趣的索引和讨论等等。

社团分类主要基于图形分割和分级聚类形成的分类算法,并在此基础上提出了很多针对具体特征形成的各种扩展分类算法。

根据图形分割原理^[3],基于 Kernighan—Lin 算法的分类采用贪婪策略,利用社团内部及外部连边最优化准则对社团进行分类,基于图的 Laplace 矩阵特征向量的谱平分类法则是利用图的 Laplace 矩阵的第二小特征值将社团分类。对于算法本身,图的精确分割是一个 NP 难题,而且社团分类问题本身分类个数及社区的顶点数未知,因此借助图形分割对社团分类仍需要大量的研究。

图的聚类是把具有相同或相似属性的节点聚合到一起,因此根据各个节点之间的相似性或强度进行社团分类。根据向网络中添加相似性最高的边或移除网络中相似性最低的边可将分级聚类分为凝聚法和分裂法。Newman 快速算法的分类是基于贪婪策略的凝聚算法,而 GN 算法是一种不断从网络中移除边数最大边的分裂算法,层次聚类算法则利用社团的层次结构特点进行分类。

除此之外,派系过滤算法根据社团彼此重叠相互关联的特点进行分类,而且各种算法相互结合,提出了基于连边密度、介数、信息中心度、随机行走、相异性或优化目标函数等等的分类算法。

2.2 BP 神经网络

神经网络是借助生物神经系统模拟神经元的特性而产生的一种抽象数学模型,利用神经网络可实

现模式识别、智能控制等复杂的功能^[4]。本文利用 BP 网络在模式识别方面的特性实现社团分类。

BP 网络作为一种基于误差反向传播算法的人工神经网络,具有较强的分类、模式识别及知识表达能力^[5]。BP 网络一般是具有输入层、隐层和输出层的三层前馈网络。通过正向和反向学习过程,调节各层神经元连接权值得到理想输出。正向传播时,输入信息通过输入层经加权到隐层再加权到输出层,若输出层与理想输出不符则反向传回网络,将误差信息返回修改各层神经元权值,此过程逐渐迭代直到得到允许误差范围内的理想输出。

BP 网络在线性与非线性可分性应用中均有不错的效果,由于社团之间的关系因连边不同会发生变化,BP 网络通过权值调整模拟这种微妙变化,所以社团分类本身的特点可以利用 BP 网络进行求解,与 BP 网络求解过程相符。

3 成员关系网络设计与求解过程

由 BP 网络原理可知,对社团分类由信息的正向传播和误差的反向传播两个过程组成。对社团的关系通过权值不断迭代,不断调整,经过学习训练之后达到社团分类的效果。空手道俱乐部内部成员网络^[6]如图 1 所示。

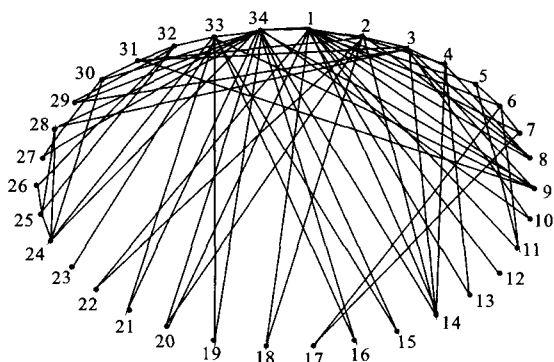


图 1 Zachary 研究的空手道俱乐部内部成员关系图

对于成员关系的拓扑结构^[7],定义一个成员为一个节点,节点之间的连接关系用邻接矩阵表示。定义如下:

$$a_{ij} = \begin{cases} 1, & \text{若节点 } i \text{ 与节点 } j \text{ 邻接} \\ 0, & \text{若节点 } i \text{ 与节点 } j \text{ 不邻接} \end{cases}$$

通过分析图 1 的关系数据,根据节点个数及分类关系,将社团分类的 BP 神经网络模型结构设计如图 2 所示。

图中 X_k 表示输入层,输入空手道成员关系的邻接矩阵向量值, H_j 表示隐层,单元个数通过实验确定, Y_i 表示输出层,输出为每一组输入对应的社团

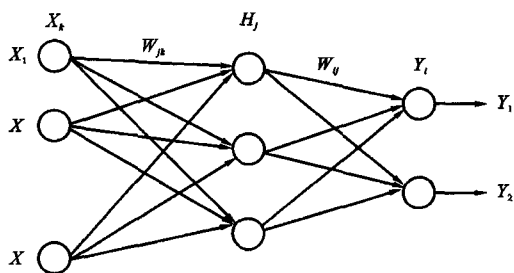


图2 BP神经网络结构设计图

分类。

本模型中隐层个数经实验确定取3个足够,活化函数采用 sigmoid 函数:

$$\varphi(x) = \frac{1}{1 + \exp(-\beta x)}$$

令 s 表示已知样本的标号,则 $s = 1, 2, \dots, n, n < 34$ 。本实例中,当第 s 个样本输入 BP 网络时,记输入单元取值为

$$I_k^s = X_k = [x_1(k), x_2(k), \dots, x_{34}(k)];$$

隐单元记为 $H_j^s (j = 1, 2, 3)$;

输出单元记为

$$O_i^s = Y_i = [y_1(i), y_2(i)]; \text{ (其中 } i = 1, 2)$$

如果给定 W_{ij} 和 W_{jk} , 对于任何输入, 网络中所有单元取值可以确定。对于样本 s , 隐单元的输入记为 h_j^s , 则输出为

$$\varphi(h_j^s) = \varphi\left(\sum_{k=1}^n W_{jk} I_k^s\right) \quad (1)$$

输出单元的输入记为 O_i^s , 将式(1)代入则为

$$o_i^s = \sum_{j=1}^3 W_{ij} H_j^s = \sum_{j=1}^3 W_{ij} \varphi\left(\sum_{k=1}^n W_{jk} I_k^s\right) \quad (2)$$

将式(2)代入最终输出, 则为

$$O_i^s = \varphi(o_i^s) = \varphi\left(\sum_{j=1}^3 W_{ij} \varphi\left(\sum_{k=1}^n W_{jk} I_k^s\right)\right) \quad (3)$$

若取得合适的权值 $\{W_{ij}, W_{jk}\}$, 则对于任一社团学习样本的输入 I_k^s , 社团理想输出即为确定的(1,0)或(0,1), 分类就可以确定。

对于 s 的理想输出设为 R_i^s , 若社团分类尽可能正确, 只要 R_i^s 与 O_i^s 尽可能没有差别, 采用最小二乘法作为能量函数:

$$E(W) = \frac{1}{2} \sum_{i,s} (R_i^s - O_i^s)^2 \quad (4)$$

只要选取适当的 W 值, 使得 $E(W)$ 最小或极小, 则 R_i^s 与 O_i^s 差异最小, 将式(3)代入式(4)得

$$E(W) = \frac{1}{2} \sum_{i,s} [R_i^s - \varphi\left(\sum_{j=1}^3 W_{ij} \varphi\left(\sum_{k=1}^n W_{jk} I_k^s\right)\right)]^2 \quad (5)$$

对于变量 W_{ij} 和 W_{jk} , $E(W)$ 为一个连续可微的

非线性函数, 采用梯度下降法求权极小值, 通过迭代运算, 求出 $E(W)$ 的局部极小, 对 W_{ij} 和 W_{jk} 而言, 梯度下降时采用 δ -法则^[5], 极大降低算法的工作量。具体过程参考相关 BP 算法推导。

4 分类设计与仿真结果

本实例中, 根据邻接矩阵成员之间关系的定义, 图1的网络拓扑结构由成员间的相互关系即可表示矩阵 A 如下:

$$A = \begin{matrix} & \begin{matrix} x_1 & x_2 & x_3 & & & & x_{34} \end{matrix} \\ \begin{matrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{34} \end{matrix} & \begin{bmatrix} 001 & \cdots & 0 \\ 001 & \cdots & 0 \\ 110 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 000 & \cdots & 0 \end{bmatrix} \end{matrix}$$

对于输入向量, 每一个行向量代表一个成员与其他成员的关系, 特别的, 成员自身的关系不考虑, 其关系值为0。将34个成员分为AB两组实验数据, A组26个作为已知样本, 8个作为检验样本, B组30个作为已知样本, 4个作为检验样本, 每组检验时分两种情况, A1组随机抽取检验样本, A2、B2组选取与两个社团均有密切关系的作为检验样本, B1组选取与两社团均有密切关系且在两个社团中连边较多的节点作为检验样本, 输出以两个分类为例。

实验设计中将矩阵的数值看作样本向量的属性值。在选择样本时不能存在孤点, 即向量值不能为0, 否则重新选择样本。

对于实验中BP网络的模型参数^[5], 经过几个组合的反复测试, 考虑到下降和学习速度以及收敛和精度要求, 设置如下: 学习步长为0.01, 误差为 $1e-10$ 。样本以每个节点对应的行向量作为数据来源。

根据以上参数设置与BP学习算法, 对已知样本进行学习, 得到四组样本学习曲线如图3所示。

然后利用检验样本进行检验, 检验结果如图4所示。

从图中可以看出, 对于给定的学习步长和精度条件, 四组样本均有很好的收敛效果, 两个曲线各自的误差函数数值也说明社团分类的正确性。

虽然实验条件和设计过程与实际环境不完全一致, 但从图中可以看出, 某些节点同时属于两个社团, 干扰性比较大, 而且如果选择的节点相互之间连边密切或者同时去除度比较大的节点, 会对网络有大的干扰, 以上四组数据均使用于干扰性大的节点作为检验样本, 仍能正确进行分类, 说明本文使用的方

法是可行有效的。

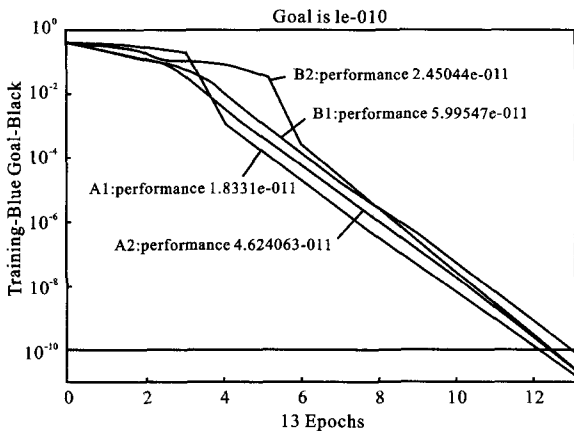


图3 实验样本学习曲线

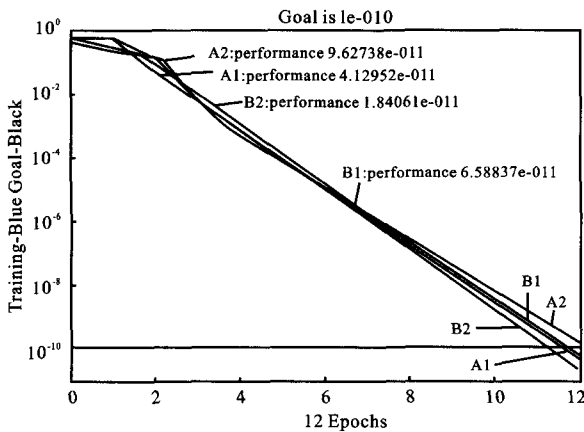


图4 实验样本检验曲线

5 结束语

本文通过分析复杂网络的社团结构,结合 BP 神经网络在模式识别方面的特点,利用二者关系与实现上的相似性以及节点之间关系的动态相关性,对社团分类进行计算机模拟分类,根据成员之间邻接矩阵的关系定义,简便而有效地实现了 BP 神经网络对复杂网络社团的正确分类。

在设计和实验中虽然达到了目标,但还存在许多不足:邻接矩阵表示成员关系是一种二值分类,数据的矩阵值不能完全代表节点在社团中的属性;实验是在已知分类结果条件下以 2 个输出为例进行模拟,样本的统计特性及结果的对比性还需要进一步

加强。

虽然本文是在各种限制条件下对复杂网络的社团进行分类,但这种应用的结果对复杂网络的研究具有积极意义和参考价值。社团分类是复杂网络研究的一个重要特征,对社团分类的研究需要在不同领域通过不同途径进行测试,建立符合实际复杂网络的模型,接下来对此问题作者及研究小组将进一步深化本文的工作;通过社团的其他属性进行分类,比如边的权值,在邻接矩阵中将节点的中心度体现出来;然后通过进行多层次的多个分类进行结果比较,同时通过不同神经网络模型对不同社团进行分类,从横向和纵向的结果进行比较,使结果更有说服力,更符合实际网络模型。

参考文献:

- [1] 汪小帆,李翔,陈关荣. 复杂网络理论及其应用[M]. 北京:清华大学出版社,2006:162-193.
- [2] Newman M E J. Finding community structure in networks using the eigenvectors of matrices[J]. Phys Rev E, 2006(74):36-104.
- [3] Shen H W, Cheng X Q, Guo J F. Quantifying and identifying the overlapping community structure in networks [J]. journal of statistical mechanics—theory and experiment, 2009:7-42.
- [4] 方向,冯烟利. 基于人工神经网络的分簇多播路由算法[J]. 微电子学与计算机,2010,27(5):78-81.
- [5] 杨俊,陈贤富. 基于 KPCA 和 RBF 网络的文本分类研究[J]. 微电子学与计算机,2010,27(3):123-125.
- [6] Zachary W W. An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research,1977,33:452-473.
- [7] 何大韧,刘宗华,汪秉宏. 复杂系统与复杂网络[M]. 北京:高等教育出版社,2009:107-145.

作者简介:

张兰华 男,(1979-),博士,讲师.研究方向数据库优化和复杂网络建模算法与仿真.

孙岩 女,(1974-),博士,副教授.研究方向为人工智能和图像处理.

薛绍伟 男,(1973-),博士.研究方向为复杂网络算法.

唐一源 男,(1966-),教授,博导.研究方向为神经信息学.