第33卷 第12期 计 算 机 工 程 2007年6月
Vol.33 No.12 Computer Engineering June 2007
·博士论文· 文章编号：1000—3428(2007)12—0016—03 文献标识码：A 中图分类号：TP312

基于网络社区模块结构的特征选择性能评价

岳 训[1,2]，迟忠先[1]，莫宏伟[3]，郝艳友[1]
(1. 大连理工大学计算机科学工程系，大连 116024；2. 山东农业大学信息科学与工程学院，泰安 271018；
3. 哈尔滨工程大学自动化学院，哈尔滨 150001)

摘 要：利用网络社区模块结构作为特征选择的度量指标，给出了一种基于全局拓扑结构的特征选择性能评价方法。对一种基于免疫学原理的数据压缩和特征提取模型——人工免疫网络进行了验证，通过对数据特征提取前的抗原数据网络和特征提取后的记忆网络的网络社区模块结构的对比，达到对人工免疫网络(aiNET)的特征提取性能评价的目的。实验结果证实了人工免疫网络模型可以保持网络拓扑结构上的稳定性，验证了利用网络社区结构作为特征选择度量的合理性。

关键词：特征选择性能评价；网络社区结构；人工免疫网络

Feature Selection Measurement Approach Based on Community Modularity Structure

YUE Xun[1,2], CHI Zhongxian[1], MO Hongwei[3], HAO Yanyou[1]
(1. Department of Computer Science & Engineering, Dalian University of Technology, Dalian 116024; 2. College of Information Sciences & Engineering, Shandong Agricultural University, Taian 271018; 3. Automation College, Harbin Engineering University, Harbin 150001)

【Abstract】Taking community structure as measurement index for feature selection, a new feature selection measure approach based on modularity coefficient community structure is proposed. Artificial immune network is a type of competitive learning algorithm which is capable of extracting relevant features contained in dataset. It uses the "internal image" memory network to eliminate data redundancy and feature extraction. The new approach is used to analyze the community structure between the input pattern (antigen) and memory network. Experiment study shows the newly approach is proved reasonable and viable.

【Key words】Feature selection measurement; Community structure; Artificial immune network

目前，常用的特征选择度量技术有：用学习算法的准确率来评估特征子集优劣的准确性度量，用样本集中不一致样本率来评估的一致性度量和与分类器无关的筛选器评价方法等[1,2]。本文借助于将信息领域中的熵函数作为评价测度技术的启发[3]，从特征选择模式的全局拓扑结构特征考虑，用网络结构描述特征选择模式，运用一种体现网络拓扑结构的指标——网络社区模块结构作为特征选择的度量，给出了一种基于全局拓扑结构的特征选择性能评价新方法。

1 基于网络社区结构的数据特征选择性能评价技术

1.1 构建网络模型

定义 1 对于数据集 $X=\{x_1,x_2,...,x_n\}$，则 X 的网络结构用图 G (V,E) 来描述。其中节点V代表为用来代表输入数据集中不同的个体，边 E 为数据个体之间的相关程度，对于任意 $x_i,x_j\in X$ 、 $x_i\in R$，两个点的距离 $d(x_{i1},x_j)$ 如果小于给定阈值 Nst，则这两点间存在一条边，图中的边可以表示为连接图形式，由这些边和相应的顶点组成 X 的数据网络。其中，两个点的距离 $d(x_{i1},x_j)$ 计算要根据不同数据编码选择相应的距离测度方法，最为常用的一些重要距离测度有欧氏距离、S阶Minkowski测度、Chebychev距离、平方距离、非线性测量等。

1.2 网络社区结构的定量描述：模块系数 Q

网络社区结构最早是由Girvan and Newman于 2002 年提出的[4]，下面给出网络社区结构模块系数Q(modularity coefficient Q)的定量描述方法。

定义 2 假设有一任意网络，其内有 Nc 个社区结构，网络可表述为 Nc×Nc 矩阵，矩阵中的每个元素 e_{ij} 表示从社区 i 的节点开始连接社区 j 中的节点的边 r 的数量占全部的概率，e_{ii} 表示从社区 i 内部连接边的数量占全部的概率，Nc×Nc 矩阵中任一行(列)之和，$a_{i,}=\sum_j e_{i,j}$ 表示连接社区 i 中的节点的边的数量占全部的概率，网络社区结构的模块系数 Q：定量描述方法可定义为

$$Q=\sum_I (eii-a_i^2) \tag{1}$$

推论 若网络是一个具有 nc 个社区，节点个数相同的完全图结构，并且社区之间没有连接，则 Q 取值为 1− 1/nc。

证明

$Q=\sum_I (e_{ii}-a_i^2)$ ：由于社区之间没有连接，$a_i=e_{ii}$

$=\sum_I (e_{ii}-e_{ii}^2)$

$=\sum_I e_{ii}-\sum_I e_{ii}^2$ ：由于 nc 个社区为完全图结构，则

$e_{ii}=1/n_c$； $\sum_I e_{ii}=1$

基金项目：国家自然科学基金资助项目(60305007)
作者简介：岳 训(1968－)，男，副教授、博士生，主研方向：数据挖掘，人工免疫算法；迟忠先，教授、博导；莫宏伟，副教授、博士；郝艳友，工程师、博士生
收稿日期：2006-06-27 **E-mail：**ywy123@dl.cn

$=1-1/nc$

nc 越大，Q 取值越趋向于 1。

Q 的值域为–1<Q<1，Q 值越大，表明网络中网络社区结构的划分越精确，即网络中属于同一社区的节点尽可能密集连接，网络中不同社区之间有节点，则尽可能松散连接；当网络中不存在网络社区结构，则 Q = 0，如标准的随机网络。当网络中存在很差的网络社区结构时，网络中节点各自为营时(自己属于自己的社区)，社区内将没有内部的边，Q 取负值，一般来说，Q 取值在 0.3 ~ 0.7 之间。

2 人工免疫网络模型(aiNET)

人工免疫网络模型是通过借鉴和利用生物免疫系统的性质和机制开发用于解决工程和科学问题的智能系统技术，它已应用在数据挖掘中的特征抽取、快速进化、学习和记忆特性等领域。Timmis和de Castro[5]两人分别提出了可用于数据压缩和聚类的免疫网络智能信息处理模型(RLAIS和aiNet)，aiNet的主要功能就是运用基于免疫网络亚动力学思想，最终用一个小规模的“内镜像”记忆网络映射源输入数据集，从而达到将数据压缩的目的，再析取出源数据集中的特征。aiNet(artificial immune network)数据数特征提取模型算法基本思路如下：

Step1 系统参数初始化；

Step2 初始化网络状态，免疫网络处于一种平衡状态，抗原输入；

Step3 针对每一个抗原；

Step3-1 根据抗原与免疫网络的抗体的亲合程序进行进化操作，如选择、变异、交叉等算子操作，导致免疫网络的扩展；

Step3-2 与免疫网络内其它的抗体相匹配导致免疫网络的压缩；

Step4 对个体进行免疫调节，自适应变异算子，产生新的个体，加入免疫网络内，转入 Step3；

Step5 最后得到的免疫网络称为“内镜像”记忆网络，它是输入抗原集的压缩结果，析取出其特征。

3 实验及性能分析

为了达到对人工免疫网络的特征提取性能评价的目的，重点研究数据压缩前的抗原网络和压缩后所提取的记忆网络的网络结构对比分析。

3.1 数据集和运行初始参数的设置

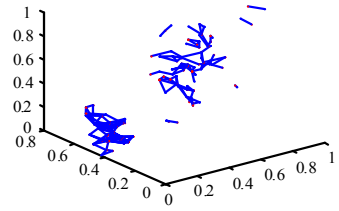
所用测试数据集选取机器学习领域常用的 UCI 中的 Iris 标准数据集和 Statistical Innovations Inc 的 cancer 数据集，测试平台为 1.6GHz Intel Pentium4 CPU，256MB 主存。由于 aiNET 数据特征模式的效果主要取决于 aiNet 运行时设置的运行初始参数，为了保证 aiNet 算法的运行，表 1 为两个数据集所采用的经验参数值。

表 1 aiNet 算法运行参数

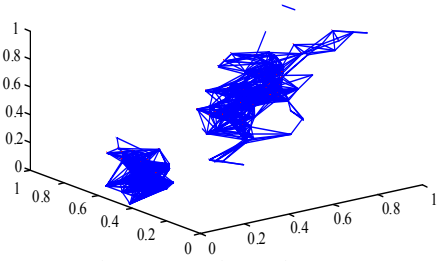
编号	变量	Iris	cancer	含义
1	n	4	4	每个抗原与抗体匹配后选取的数目
2	N	10	10	克隆数目
3	qi	0.2	0.2	克隆后选择的数目
4	gen	40	40	循环代数
5	ts	0.15	0.56	压缩比例
6	mi	4	4	变异比例
7	Sc	0.01	0.01	运行终止阈值

3.2 网络构建

根据数据实数编码，选择欧氏距离测度方法，图 1 是抗原数据网络 Iris 的可视化表示，由于抗原网络 Iris 为四维数据，图 1 仅为选取其中三维的可视图。从图 1(a)、图 1(b)的对比中可明显看出，抗原数据网络 Iris 中含有两个社区的结构特征。



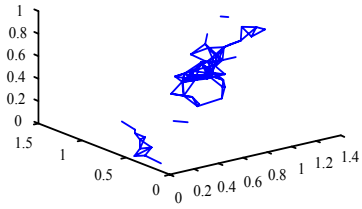
(a) Nst=0.2, edge number=243



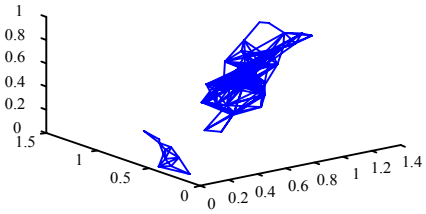
(b) Nst=0.64, edge number= 2620

图 1 抗原数据网络 Iris 的网络可视图

图 2 是经 aiNET 压缩后所形成的记忆网络 Memory-iris 的网络可视化表示。



(a)Nst=0.62, edge number=123



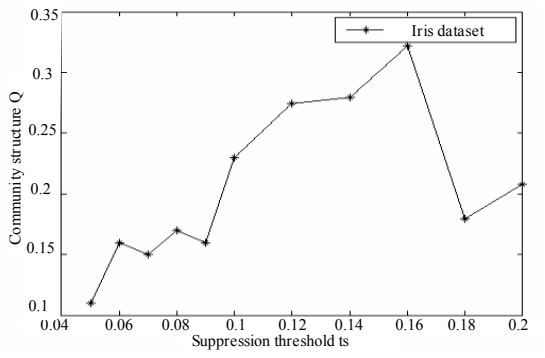
(b)Nst=0.81, edge number= 676

图 2 抗原数据网络 aiNET 压缩后所形成的记忆网络 Memory-iris 的网络可视图

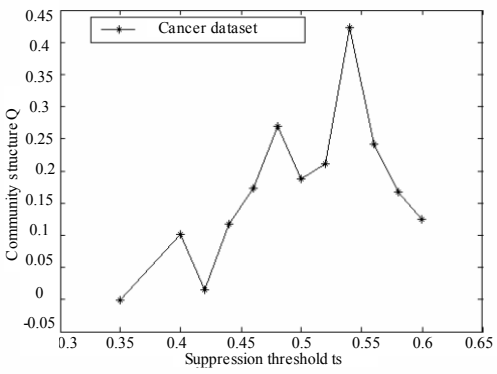
从图 1 和图 2 的对比中容易看出，经 aiNET 压缩后所形成的记忆网络 Menory-iris 保持了数据压缩前的抗原数据网络 Iris 的网络拓扑结构。必须指出，阈值 Nst 是影响网络规模的关键参数，在后面的研究中，我们选取模式数据集各自的平均距离为阈值 Nst 的值，其中，抗原网络 Iris 为 0.640 3、cancer 为 0.678 2，相应的记忆网络 Menory-iris 为 0.611 7、Menory-cancer 为 0.637 1。

3.3 对人工免疫网络的特征提取性能评价

由于 aiNET 模型的特征提取效果主要取决于 aiNet 运行时设置的参数，其中压缩比例 ts 是最主要的因素，目前，由于缺乏能对人工免疫网络的压缩性能评价方法，为了保证 aiNet 算法的运行，不同数据集所采用的都是经验参数值选取。图 3 是压缩比例 ts 选取不同值时模块系数 Q 的变化情况，图 3(a)中，针对 iris 数据集，当 ts=0.16 时，模块系数 Q 有最大值，意味着这时得到的 iris 数据集压缩后所形成的记忆网络具有好的网络社区结构，从而能保持较好的原数据集的网络拓扑结构；图 3(b)中，针对 cancer 数据集，当 ts=0.57 时，得到的记忆网络能保持较好的原数据集的网络拓扑结构。



(a)iris 数据集



(b)cancer 数据集

图 3 压缩比例 ts 选取不同值时模块系数 Q 的变化情况

下面给出用模块系数 Q 作为对人工免疫网络模型算法的压缩性能评价指标的实例。

从理论上分析，经 aiNET 压缩后所形成的记忆网络在网络社区结构上要保持着与数据压缩前的抗原数据网络相同的特征，通过对 aiNET 压缩前后的网络社区结构的对比，模块系数 Q 越接近，则说明人工免疫网络模型算法的压缩性能越好，将其相近程度作为对人工免疫网络(aiNET)的性能进行评价的量化指标：

$$\eta = \frac{Q - Q'}{Q} \tag{2}$$

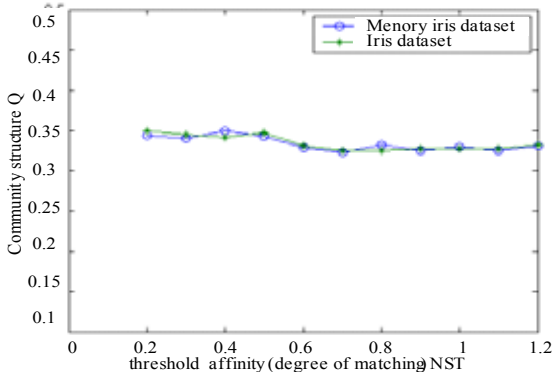
其中，Q 和 Q' 分别是压缩前的抗原数据网络和经 aiNET 压缩后所形成的记忆网络的模块系数。

表 2 给出针对本文两组实际数据集运行时的不同初始参数值，由于 aiNet 运行时需要设置运行初始参数较多，表 2 只给出较关键的 3 个参数组合时情况，实验数据与实际效果表明本文提出的评价量化指标的有效性。

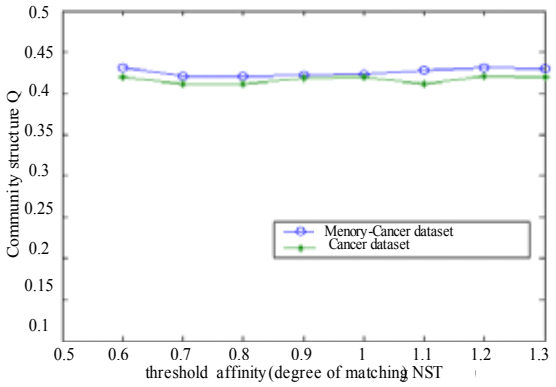
表 2 iris 数据集在不同初始参数值运行时的评价量化指标 η

Data set	qi	ts	sc	η /%
Menory-iris (1)	0.1	0.001	0.01	21.5
Menory-iris (2)	0.15	0.01	0.1	13.8
* Menory-iris (3)	0.2	0.15	0.01	3.6
Menory-iris (4)	0.25	0.15	0.1	5.6
Menory-iris (5)	0.3	0.2	0.01	23.2

图 4 是网络形成时距离阈值 NST 选取不同值时模块系数 Q 的变化情况。图 4(a)中，针对 iris 数据集，当 ts=0.16 时，模块系数 Q=0.34±0.03，这时得到的 iris 数据集压缩后所形成的记忆网络具有与原数据集的网络拓扑结构相近的模块系数 Q，从而能保持较好的压缩效果；图 4(b)中，针对 cancer 数据集，当 ts=0.57，模块系数 Q=0.43±0.2，也能保持较好的压缩效果。



(a)iris 数据集，ts=0.16



(b)cancer 数据集，ts=0.57

图 4 网络形成时距离阈值 NST 选取不同值时模块系数 Q 变化情况

4 结论及未来工作

本文给出了一种基于全局拓扑结构的特征选择性能评价新方法，对一种基于免疫学原理的数据压缩和特征提取模型——人工免疫网络进行了验证，通过对数据特征提取前的抗原数据网络和特征提取后的记忆网络的网络社区结构系数的对比，达到对人工免疫网络的特征提取性能评价的目的。

由于 aiNET 数据压缩和聚类效果主要取决于 aiNet 运行时设置的运行初始参数，通过对本文提出的模块系数 Q，对于保证 aiNet 算法的运行效果，特别是设和选取算法运行时的参数起到量化评价作用。实验结果验证了利用网络社区结构作为特征选择度量的合理性。

但是，由于提取模式特征的拓扑结构特性复杂性差异较大，本文所给出的技术在应用领域是有限的，影响因素主要有数据类型、问题规模、样本数量等，多种特征提取方法的融合将是特征提取的发展方向。

参考文献

1 边肇祺，张学工．模式识别[M]．2 版．北京：清华大学出版社，2001.

2 Bassat M. Pattern Recognition and Reduction of Dimensionality[Z]// Krishnaiah P R, Kanal L N. Handbook of Statistics. 1982.

3 Torkkola K. Nonlinear Feature Transforms Using Maximum Mutual Information[C]//Proceedings of the International Joint Conference on Neural Networks. 2001.

4 Girvan M, Newman M E J. Community Structure in Social and Biological Networks[C]//Proceedings of the National Academy of Science, USA. 2002.

5 L N de Castro, Timmis J. Artificial Immune Systems as a Novel Soft Computing Paradigm[J]. Soft Computing, 2003, 7(8): 526-544.