

基于核聚类的近邻人工免疫网络算法研究^{*}

李 昕, 李立君, 高自成

(中南林业科技大学 机电工程学院, 长沙 410004)

摘 要: 针对经典人工免疫网络(aiNet)及改进算法中存在的运算时间长、结构复杂等问题,提出了一种改进的核聚类近邻人工免疫网络算法(KN-aiNet)。算法在aiNet的改进算法——近邻aiNet结构的基础上,以抗体数据为核心利用量子能级思想聚类,并重定义了生成抗体策略,采用区域生长法搜索拥挤距离,采用基于核函数的亲和度等方法来提高算法的聚类效果和降低算法的运算时间。聚类实验结果表明,KN-aiNet算法的聚类准确率较经典aiNet算法及近邻aiNet算法分别提高了11.53%和4.56%,而算法的运算时间较经典aiNet算法及近邻aiNet算法分别下降了0.503 s和0.823 s。

关键词: 人工免疫网络; 近邻aiNet; 核函数; 区域生长

中图分类号: TP18; TP301.6

文献标志码: A

文章编号: 1001-3695(2012)07-2464-03

doi:10.3969/j.issn.1001-3695.2012.07.017

Study in neighbor artificial immune network based on kernel clustering

LI Xin, LI Li-jun, GAO Zi-cheng

(School of Mechanical & Electrical Engineer, Center South University of Forestry & Technology, Changsha 410004, China)

Abstract: To solve the issue of high time complexity and algorithm structure of aiNet, this paper proposed a modified aiNet algorithm; KN-aiNet. Based on the structure of neighbor aiNet, the algorithm took antibody data as core, then clustered data around the core, used energy level as affinity. The algorithm redefined the creation of antibody, took region growing to match distance, took kernel function to improve clustering effect and reduce time complexity. The simulation proves the clustering accuracy of KN-aiNet improves 11.5% compared with aiNet, 4.56% compared with neighbor aiNet, and time complexity declines 0.503 s compared with aiNet, 0.823 s compared with neighbor aiNet.

Key words: aiNet; neighbor aiNet; kernel clustering; region growing

0 引言

经典人工免疫网络算法(aiNet)借鉴自然界生物体免疫系统功能,利用生物体本我非我识别的思想实现数据的聚类,具有较好的浓缩性和鲁棒性^[1],但也存在抗体网络随机指定,容易陷入局部收敛;抗体抗原采用全局匹配方式搜索,增加了网络的运算时间;网络的抑制阈值和抗体刺激阈值需根据具体情况手工给定等问题。对于这些问题,国内外研究者提出了很多改进方法,将支持向量机算法^[2]、粒子群算法^[3]、多样性算子^[4]融入aiNet以提高聚类准确性;在网络自身结构上宋晓琳等人^[5]采取矢量矩,Hamaker等人^[6]采用非欧编码为度量标准以改善抗体与抗原的距离函数;郑永煌等人^[7]采用阴性选择算法自适应确定刺激阈值和抑制阈值。但是这些改进都增加了免疫网络结构复杂度,导致免疫网络结构复杂、算法实时性降低,不利于人工免疫网络算法在工程领域的应用。

针对经典算法及其改进算法存在的问题,Gong等人^[8]提出了一种非支配近邻免疫算法,算法中根据抗体周围的拥挤程度定义克隆等级,取得了较好的结果。本文在近邻免疫算法的研究基础上,提出一种核聚类近邻人工免疫网络算法(KN-aiNet),算法中利用SVM与RLAIS算法自适应生成核心,以核函

数为度量进行聚类,从而将数据分类问题转换为核函数量子的能量聚类问题。KN-aiNet采用划分抗体分布和近邻选择的方法以保持网络的多样性,防止陷入局部收敛;利用区域生长法搜索抗体抗原的拥挤度以减少拥挤距离全局搜索的执行时间。仿真实验结果及算法对比表明,改进算法的聚类准确率和聚类速度较经典aiNet算法及近邻aiNet算法均有了一定的提高。

1 核聚类近邻 aiNet 网络算法结构

KN-aiNet算法中核的设定是借鉴于SVM^[9]算法中核函数空间距离定义与RLAIS^[10](资源受限人工免疫系统)中的ARB(人工认知球)的思想而产生的。在RLAIS中,ARB代表了同一类的B细胞中的三个属性,即匹配规则、数据项和刺激度。而SVM中的核函数空间距离定义对数据分类问题起着重要的作用。刘大有等人^[11]将核聚类的方法融入人工免疫网络中解决分类数据边缘不清晰的问题,取得了较好的效果。

1) 核心的选取

初始核心数据可以采用样本特征数据作为抗体数据进行有导师聚类,也可以按空间分布随机选取抗体作为核心数据,这种核心选取方法既保证了抗体的可靠性,又体现了抗体空间分布的特性,可提高算法的聚类效率。在聚类过程结束后按照

收稿日期: 2011-12-21; 修回日期: 2012-02-08 基金项目: 林业公益性行业科研专项基金资助项目(201104090)

作者简介: 李昕(1984-),男,湖南长沙人,博士研究生,主要研究方向为人工智能、免疫算法(lixin7059@gmail.com);李立君(1964-),女,湖南宁乡人,教授,博导,博士(后),主要研究方向为现代林业技术与装备;高自成(1968-),男,湖南长沙人,博士研究生,主要研究方向为林业机械化工程。

聚类的欧氏距离重新调整核心的位置,KN-aiNet算法中的核心初始选取式如下:

$$t_i = \frac{\sum_{k=1}^n a_k}{mp} \quad i=1,2,\dots; k=1,2,\dots,n \quad (1)$$

式(1)可看做把流行拓扑空间分为 n 块, t_i 代表每块中的数据点; $\sum_{k=1}^n a_k$ 为数据点个数之和;参数 p 的作用在于控制面积的大小以控制生成抗体的规模;参数 m 为聚类平面中分块的数量,可以通过分块的数量来保持抗体的多样性。重新定义的核心选取方式设置了参数 p 来控制抗体的规模和保持网络的全局多样性,在应用中可以根据大小需要调整。

2) 亲和度的计算

将核心抗体作为原子的核心,抗原数据作为核心周围量子,这样就把免疫网络中抗体抗原的距离转换为以抗体为中心的能量表示。为方便算法后续计算过程中能量的表示,规定能量本征态能量为 0。忽略量子间动能的影响,两个量子间的能量度量函数由两个量子间的电势能表示,取无穷远处电势能为 0,则两个距离为 r 的量子电势能表示为

$$E = eU = -e \frac{kZe}{r} = -\frac{kZe^2}{r} \quad (2)$$

由于数据聚类的数据点之间属性相同,为简化计算,规定式(2)中的量子为单位电量 $Ze=1$ 。取两个量子间电势能阈值为 σ_e ,量子间的 $E \leq \sigma_e$ 时,数据属同类,从而实现量子能级间的聚类操作。抗体与抗原的亲合力函数采用核函数表示,并将量子间的电势能代入核函数中,抗体 a 与抗原 b 之间的能级度量的核函数可以表示为

$$K(a,b) = [\varphi(a) \cdot \varphi(b)] = \|E(a-b)\| \quad (3)$$

抗原与抗体通过核函数映射到核空间后,它们的亲合力函数可以表示为

$$d(a,b) = \|\varphi(a) - \varphi(b)\| = \sqrt{|K(a,a) - 2K(a,b) + K(b,b)|} \quad (4)$$

3) 拥挤距离中的区域生长法

KN-aiNet 所定义的拥挤距离是根据目标空间中抗体与周围数据之间的亲和度判定的。拥挤距离是对抗体的近邻抗原拥挤程度的度量准则,也是反映抗体的量子能量的量度。一个抗体 d 在其种群 D 的拥挤距离定义为

$$I(d,D) = \sum_{i=1}^k \|X_i - X_j\|, d \in D \quad (5)$$

其中: X_j 为待测抗体数据; $I(d,D)$ 代表此抗体在种群 D 中的综合拥挤程度。

物理学中距离原子核越近的电子迁跃所需的能量越大。当抗体周围的抗原数量较多时,由于无法单纯使用拥挤距离反映出抗体周围抗原的能量程度,本文引入克隆等级的概念,即抗原与抗体的拥挤距离越小克隆系数越小,拥挤距离越大克隆系数越大。抗原等级与拥挤距离的设定有利于赋予抗体克隆优先权,保持抗体克隆的多样性。克隆等级定义为

$$T^r(a_1 + a_2 + \dots + a_n) = T^1(a_1) + T^2(a_2) + \dots + T^n(a_n) \quad (6)$$

其中: T^r 为克隆算子; $a_1, a_2, \dots, a_n$ 为待克隆的网络细胞; $T^1, T^2, \dots, T^n$ 为根据各网络细胞实际拥挤程度定义的单体克隆系数。

经典 aiNet 及其近邻 aiNet 算法中的搜索方式为全局搜索,极大地增加了算法的运算时间。KN-aiNet 算法中,为了提

高抗体与抗原之间的匹配速度,聚类的亲和力搜索方式采用了图像学中的区域生长法^[12]的方式,将每个核心抗体数据点看做一个量子种子点,而抗体与抗原数据点间的距离度量则转换为量子之间的核函数亲和力度量。以种子点抗体位置为中心寻找周围的抗原,若周围抗原数目大于亲和力阈值,即归为一类;再以四周抗原为种子抗体去匹配周围的数据,若小于亲和力阈值则停止搜索。依此类推,可以得到较好的聚类效果。以抗体为核心的区域生长搜索方式如式(7)所示,式中 σ_s 表示亲和力阈值。

$$d(a,b) \geq \sigma_s \quad (7)$$

抗体多样化调整后前后聚数效果比较如图 1 所示。

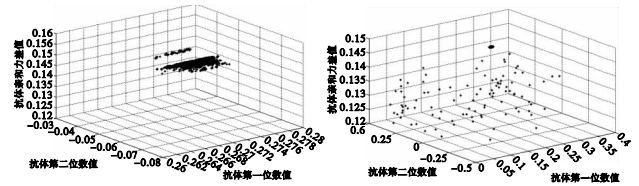


图1 抗体多样化调整后前后聚数效果

2 KN-aiNet 算法流程

抗体抗原拥挤距离和克隆系数确定后,根据多样性抗体共生、少数抗体激活的思想,选取拥挤程度较小的抗体个体作为活性抗体,根据其抗体拥挤程度选取相应的克隆系数实行克隆操作,这样有利于算法搜索网络中的稀疏区域,保持算法的多样性。拥挤距离的搜索方式采用区域生长法所确定的式(7)的搜索方法,有利于减少算法搜索的时间。KN-aiNet 的算法流程如下所述:

定义 n 为活性群体最大规模, A_i 为活性网络细胞群。

a) 初始化,按照式(1)从数据库中生成一定数量的数据点作为抗原细胞作为网络的初始输入数据。

b) 构建以抗体为核心的核聚类免疫网络。

c) 按照式(7)所确定的区域生长法计算所有输入抗原细胞与核心的亲和力,亲和力函数标准采取按式(4)计算的核空间亲和力度量。

d) 选取拥挤距离较大的前 n 个网络细胞组成活性网络细胞群 A_i 。

e) 对 A_i 中的活性网络细胞按照式(6)进行比例克隆。

f) 根据下式对克隆细胞变异,变异概率与父代亲和力成反比,并保留父代群体于网络中。

$$C' = C + \alpha N(0,1) \quad (8)$$

$$\alpha = (1/\beta) \exp(-f) \quad (9)$$

其中: C' 是细胞 C 产生变异后形成的新细胞; $N(0,1)$ 是一个均值为 0、标准偏差为 1 的高斯随机变量; β 是用于控制函数指数衰减的变量; f 是经过标准化处理后细胞适应值。

g) 计算变异后网络细胞的亲和力。

h) 选择亲和力最高的网络细胞组成新的网络,并计算新网络的细胞亲和力。

i) 计算网络中所有细胞的亲和力,如果小于阈值则亲和力高的细胞予以保留,其他细胞则抑制。

j) 引入一定百分比的随机生成细胞,返回步骤 b)。

k) 输出产生的网络细胞。

3 KN-aiNet 算法仿真分析

本文算法验证的实验硬件环境为 Intel 四核 Q8200 处理器,软件环境为 Windows 7 及 MATLAB 2010a。在同样的计算量下,不同的核心参数 p 所产生的聚类精度是不同的,以 UCI 数据集中的 Iris、Wine、Diabetes 三个集合进行测试,综合可得当 $p=0.3$ 时的识别率最高,因此选此数值为 p 的默认参数。测试结果如由图 2 所示。

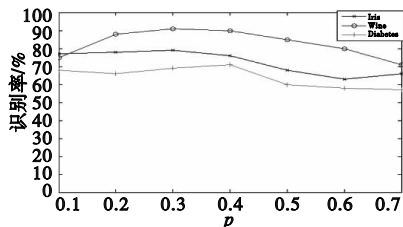


图2 核心参数 p 对识别率的影响

提取一组类螺旋形的 105×2 矩阵的数据进行聚类仿真实验,这组数据能够较好地反映数据的复杂性状态。将 KN-aiNet 算法应用于这组数据,亲和力阈值 $\sigma_s=0.01$ 和 $\sigma_s=0.03$ 两种情况下网络聚类效果如图 3 所示。从图 3 中可以看出, $\sigma_s=0.01$ 时较 $\sigma_s=0.03$ 时得到的分类数目较多、分类精度较高,由此可见数据的分类数目多少和精度大小是由 σ_s 的数值大小决定的, σ_s 数值越小则数据分类数目越多、精度越大。在实际应用中应根据需要设定 σ_s 数值。

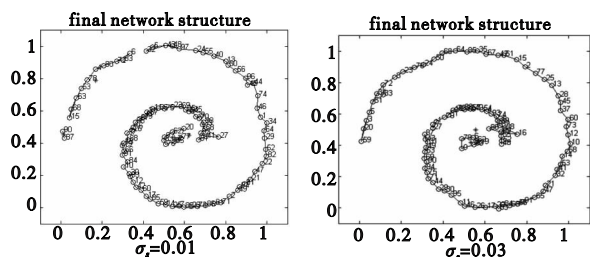


图3 KN-aiNet算法聚类效果

通过对经典人工免疫网络算法、非支配近邻 aiNet 算法与核聚类近邻 aiNet 网络算法对 UCI 数据集中 Iris 实验数据对比分析,聚类识别率与时间的定义如下:

$$\text{accuracy} = \frac{\text{acc}(\text{Ag})}{\text{all}(\text{Ag})} \times 100\% \quad (10)$$

$$\text{time} = O(\text{algorithm}) \quad (11)$$

其中: $\text{acc}(\text{Ag})$ 为正确聚类的抗原数目; $\text{all}(\text{Ag})$ 为抗原总数目; $O(\text{algorithm})$ 为各算法的运算时间。对比实验中参数是根据对 Iris 实验数据多次实验所得的最佳参数,即抗体数量 $N=10$, 变异率 $m_i=4$, 阈值 $\sigma_s=0.01$ 。分别选取无干扰样本、轻度随机干扰样本、高度随机干扰样本三类样本数据进行聚类实验。由图 4、5 可以看出,由于 KN-aiNet 算法改变了初始抗体产生规则,增加了抗体种群的多样性,因此抗体聚类的准确率较经典免疫网络算法和近邻免疫网络算法都有所提高。而且 KN-aiNet 算法将区域生长法融入了亲和力匹配算法中,运算时间有所降低。在默认的系数 $p=0.3$ 时 KN-aiNet 算法的识别率较经典算法增加了 6.2%、9.6% 和 18.8%, 平均识别率提高了 11.53%, 较近邻 aiNet 算法分别减少了 1.3%、5.7% 和 6.7%, 平均识别率提高了 4.56%。而训练时间在三种情况下较经典算法分别减少了 0.38 s、0.52 s 和 0.61 s, 平均聚类时间减少

了 0.503 s, 较近邻 aiNet 算法分别减少了 0.66 s、0.84 s 和 0.97 s, 平均聚类时间减少了 0.823 s, 取得了较好的效果。

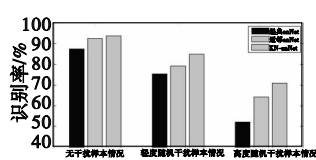


图4 KN-aiNet与aiNet、近邻aiNet识别率对比

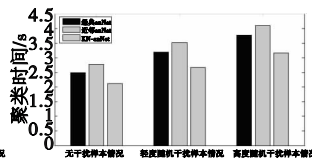


图5 KN-aiNet与aiNet、近邻aiNet运算时间对比

4 结束语

本文针对免疫网络算法的聚类特性及存在网络结构复杂、算法实时性不高的不足,采取了以下改进方法:a)构建抗体核心,并基于核心采取核函数量子能级方式聚类;b)重定义抗体生成方式以增加网络多样性;c)基于邻域 aiNet 算法中的思想,并使用区域生长法进行拥挤距离的搜索以增加算法的聚类准确率,降低聚类时间。仿真结果表明,改进算法较经典算法及其改进近邻免疫网络算法的性能和实用性有了较大的提高。

参考文献:

- [1] De CASTRO L N, Von ZUBEN F J. An evolutionary immune network for data clustering[C]//Proc of the 6th Brazilian Symposium in Neural Networks. Washington DC:IEEE Computer Society,2000:84-89.
- [2] YUAN Sheng-fa, CHU Fu-lei. Fault diagnosis based on support vector machines with parameters optimization by artificial immunization algorithm[J]. Mechanical Systems and Signal Processing,2007, 21(3):1318-1330.
- [3] LIU Li, XU Wen-bo. A cooperative artificial immune network with particle swarm behavior for multimodal function optimization[C]//Proc of IEEE World Congress on Evolutionary Computation/Computation Intelligence. 2008:1550-1555.
- [4] HART E, BERSINI H, SANTOS F. Tolerance vs intolerance: how affinity defines topology in an idiotypic network[C]//Proc of the 5th International Conference on Artificial Immune System. Berlin:Springer-Verlag,2006:109-121.
- [5] 宋晓琳,于德介,殷智宏,等. 基于矢量矩的汽车主动悬架免疫控制策略研究[J]. 湖南大学学报:自然科学版,2006,33(4):46-49.
- [6] HAMAKER J S, BOGGESS L. Non-euclidean distance measures in AIRS, an artificial immune classification system[J]. Evolutionary Computation,2004,19(1):1067-1073.
- [7] 郑永煌,李人厚. 一种改进的基于免疫网络模型(aiNet)的故障诊断算法[J]. 控制与决策,2010,12(6):847-856.
- [8] GONG Mao-guo, JIAO Li-cheng, DU Hai-feng, et al. Multi-objective immune algorithm with nondominated neighbor-based selection[J]. Evolution Computation,2008,16(2):225-255.
- [9] VAPNIK V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 2000.
- [10] THIMMS J, NEAL M. A resource limited artificial immune system for data analysis[J]. Knowledge Based Systems,2001,14(3):121-130.
- [11] 刘大有,古方明,王生生. 基于人工免疫核聚类的支持向量数据描述方法[J]. 吉林大学学报:工学版,2011,41(5):1369-1373.
- [12] LI Ming-kun, SETHI I K, LI Dong-ge, et al. Region growing using online learning[C]//Proc of International Conference on Imaging Science, Systems and Technology. 2003:73276-7279.