人工神经生长细胞结构网络在医疗诊断的应用

刘淑英[1],程国建[2],彭 方[1]
(1.咸阳师范学院 计算机系,陕西 咸阳 712000;
2.西安石油大学 计算机学院,陕西 西安 710065)

摘 要:神经网络在医疗智能诊断、细胞图像识别、信息处理等方面的应用效果显著,具有广阔的发展前景。近年来人们对自身健康的重视,人工神经网络在医疗诊断方面应用的研究,正逐渐成为人们关注的焦点。与常用的BP神经网络相比,人工神经网络生长细胞结构(GCS)网络可通过降维映射用于高维数据的聚类和可视化,其网络生成与网络结构和大小都由输入的数据加以确定。研究中采用GCS网络对患者的乳腺癌症细胞的病变类别做出了较为精确的诊断,同时进行了特性分析。并结合实际问题,对某医院的癌症测试指标数据库中699例患者数据进行试验。结果表明:该模型是一种操作简便、易于实现、性能良好的有效模型。

关键词:生长细胞结构;聚类;可视化
中图分类号:TP183 **文献标识码**:A **文章编号**:1673-629X(2009)05-0231-04

Applications of Growing Cell Structures of Artificial Neural Network for Medical Diagnosis

LIU Shu-ying[1], CHENG Guo-jian[2], PENG Fang[1]
(1. Xianyang Normal University, Xianyang 712000, China;
2. School of Computer Science, Xi'an Shiyou University, Xi'an 710065, China)

Abstract: Neural networks are significant in the application of medical diagnosis, cell image recognition, information processing and so on, with broad prospects for development. As people in recent years focus the importance of their health, artificial neural network applications in medical diagnostic research is gradually becoming the focus of attention. In comparison with BP model, growing cell structures of artificial neural network can be used in high-dimensional mapping data clustering and visualization with drop-dimensional mapping, and generation network is determined by the network structure and size of the input data. In this paper, the growing cell structures of artificial neural network is proposed for the processing parameter, it makes a more accurate diagnosis to breast cancer cells of the type of lesion and analyzes the characteristics on the network. Simulation on the testing data about the mammary cancer of the 699 sufferer shows that this algorithm is practicable and effective for actual reservoir modeling.

Key words: growing cell structures; clustering; visualization

0 引 言

随着社会的进步和人们物质文化生活水平的提高,全民健康意识也不断在加强,因此医疗诊断方面的研究备受人们关注。医学上常用的诊断方法受主观因素影响比较大,诊断结果的正误与医生医疗水平密切相关。而常用医学统计方法如线性回归模型、logistic回归模型等往往要求数据服从正态分布或经变量变换后服从正态分布,且自变量、因变量之间的关系为线性关系,但实际数据大都不能满足这些条件。人工神经网络是一种非线性动力学系统,具有自动学习和识别变量间关系的能力,对资料类型限制性较小[1,2]。人工神经网络将大量的样本病例,通过训练和学习得到网络各种神经元之间的连接权,从样本中自动获取规则,建立网络模型,并将这种模型用于对新病例的判别。许多研究都显示人工神经网络应用于临床诊断有较高的敏感性和特异性[3]。

近年来,国内外研究学者利用人工神经网络解决了很多医疗问题,据资料获悉人工神经网络理论的发展为医疗智能诊断系统提供了一条新的有效途径[4]。

收稿日期:2008-08-28
基金项目:国家自然科学基金资助项目(40572082);咸阳师范学院院内基金资助项目(07XSYK261)
作者简介:刘淑英(1982-),女,硕士,助教,研究方向为智能计算;程国建,副教授,获德国图宾根大学理学博士学位,研究方向为神经网络与进化计算、油藏模拟等。

基于这一思路，人们建立了人工神经网络式的医疗智能诊断系统，可以将医疗智能诊断系统提高到一个新的水平[5]。

但通过分析前人的研究工作不难看出，目前所采用的网络大部分是 BP 神经网络模型[6]。虽然该网络在一定条件下可以取得满意的应用效果，但还存在如下问题：

1)BP 网络是监督型学习，不能自学。它的学习状态和工作状态是截然分开的。

2)学习过程中神经网络的所有权系数都要细加调整。这就会降低学习速度。

3)BP 网络采用最优化算法的快速下降法，极有可能陷入均方误差的局部极小点，造成错误的分类结果。

而文中研究的人工神经网络生长细胞结构(GCS)模型克服了这些不足[7]。GCS 可产生降维的映射可用于高维数据的可视化或聚类。生成拓扑的基本构造单元是一个 K 维超立方体(K 必提前给定)。网络拓扑不必事先给定，而在构建网络过程中神经及其连接权值在不断增长。

GCS 模型中的学习参数是不随时间变化的常数，这就使增长过程可一直持续，直到达到一个给定大小的网络或某一问题域的性能基准得以满足为止。输入数据直接指导新神经元的生长。因此，通常会产生比预定义拓扑大小更好地反映给定输入分布的网络结构。基于上述问题，引入 GCS 网络对采样的患者数据信息进行研究与诊断。

文中主要研究如何使用 GCS 网络，根据病理信息中的大量数据进行疾病的分析和诊断。

1 原理与方法

文中利用神经网络的一种生长细胞结构模型——GCS 神经网络模型解决相关的医疗诊断问题。它可以有效地处理不确定和模糊的输入数据。

1.1 GCS 模型

GCS 模型是 Fritzke 提出的一种生长模型，GCS 可产生降维的映射用于高维数据的可视化或聚类[8]。GCS 的目标是从高维输入空间 R_n 到一个低维 k 的拓扑结构 H 产生一个拓扑保持映射。拓扑保持含义如下：

(1) 在 R_n 中靠近的输入矢量映射到 H 中相邻的节点上。

(2) H 中相邻的节点应当有相似的输入矢量映射到其上。

GCS 模型由一组节点(或神经元) 组成(节点集合记为 H)，每个节点有一个关联的 n 维参考矢量 w_c，代表其在输入空间 R_n 中的接受域中心。一个给定的节点集合及其参考矢量定义输入空间的一个特定划分，即所谓的 Voronoi 棋盘。每个节点 C 的接收域就是其参考矢量 w_c 的 Vor 区域 $v(c)$，可表示为：

$$v(c)=\{p\in R_n \mid (\|p-W_c\|<\|p-W_d\|),\ \forall d\in H, d\neq c\} \tag{1}$$

因此神经元 C 的邻域是由 R_n 的这样一些点组成，目前，所有的参考矢量离 w_c 均为最近 。

1.2 GCS 模型学习算法

GCS 模型中的学习大体上有两个阶段，即参考矢量的适配和系统的插入与连接。下是 GCS 学习算法[7]：

1)以一个 k 维网络为初始化网络结构。初始化网络结构中的神经元为($k+1$)，它是 R_n 中的随机矢量。

2) 加载一个来自输入矢量概率分布空间 $p(\zeta)$ 的一个输入信号 ζ。

3) 确定获胜神经元(具有最近参考矢量的元)。

$$\|W_{win}-\zeta\|\leqslant\|W_t-\zeta\|,(\forall c\in H) \tag{2}$$

4) 累加输入信号与获胜神经元之平方距离到一个局部误差变量 ε_{win}。

$$\Delta\varepsilon_{win}=\|W_{win}-\zeta\|^2 \tag{3}$$

5) 朝向输入信号 ζ 移动的获胜神经元及其拓扑邻元，移动幅度分别为总距离的某个百分比(移动因子)

$$\Delta\varepsilon_{win}=\eta_{win}(\zeta-w_{win}),\Delta w_l=\eta_{win}(\zeta-w_l),(\forall l\in N_{win}) \tag{4}$$

6) 若截止目前所产生的输入信号数目为学习参数 λ 的整数倍，则按下列方式插入一个新神经元：

a. 确定具有 max 积累误差的元 v_{max}，且 $\varepsilon_{max}\geqslant\varepsilon_c$，$(\forall c\in H)$

b. 通过断开 v_{max} 的最长边而生长一个新神经元，即新神经元插入在边 v_{max} 与 v_{mdn} 的中点

c. 内插参数矢量(来自 v_{max} 和 v_{mdn})，即 $w_{new}=\frac{1}{2}(w_{max}+w_{mdn})$

d. 降低 v_{mdn} 全部的邻域神经元之间的误差，即 $\Delta\varepsilon_i=-\frac{\alpha}{|N_{new}|}\varepsilon_i,(\forall i\in N_{new})$

e. 设新元 v_{new} 的误差变量为其邻元之平均值，即 $\varepsilon_{new}=\frac{1}{|N_{new}|}\sum_{i\in N_{new}}\varepsilon_i$

7)降低全部神经元的误差变量(以防溢出)。

8)若某个终止标准未满足(如最大学习步)则转第 2)步，否则终止。

2 实例测试

本节将前面讨论的 GCS 模型应用于乳腺癌症数

据库的诊断中。鉴于此病例属于常见的癌症案例，测试数据的选集比较容易且具有代表性，对模型的测试更有说服力。

针对癌症测试指标数据库中 699 例患者数据的任意 15 例取样。在实验中，抽取了 525 例应用于训练，其余的 174 例应用于测试。其中 X1 - X9 为测试指标，分别代表了患者所取样细胞的 9 种细胞性状测试指标，例如 X1 代表取样细胞厚度(thickness)；X2 代表细胞大小；X3 代表细胞形状……通过 X1 - X9 的综合测试评定，可准确诊断出乳腺肿瘤的类型，即良性(Benign)和恶性(Malignant)。

具体的测试过程：首先，新创建一个网络模型，此步骤将丢弃早先的任何网络和训练数据，并且创建一个新的初始网络。其次，加载数据，将采集好的 525 例训练数据加载到模型中。然后，根据经验值设置网络参数，此处设置增加的神经元节点为 12，每一个神经元节点的迭代步数为 5，训练数据为 525，测试数据为 174。接着，根据上面的设置训练网络，通过一个活动图表体现训练的过程，直到训练完成。当每一次有新的节点添加到网络中时，图表将会被更新。最终训练好的网络如图 1 所示。从图 1 可知，网络结构已经被更新并确定，通过其可以反映训练后的新网络的大小。注：文中数据都是经过对数化处理后得到，所有数据均为无量纲化数据。所有图例中，X 轴表示输入数据，Y 轴表示输出数据。在下面图例中，刻度单位不再给出。

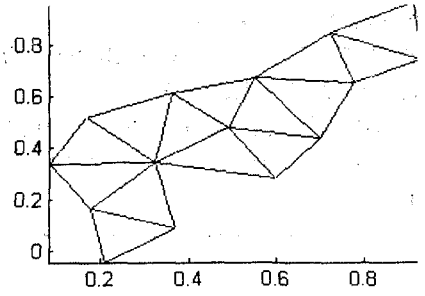

图 1　训练后的网络

最后，对测试数据库中的数据进行测试得到图 2。其中，左图为良性诊断结果，右图为恶性诊断结果。从图 2 可以得到所有分类结果(即 Benign 和 Malignant)的估计，并可以进行比较。从图中可以看到明显的两个分类区域，而且两个图形的分类结果完全吻合。

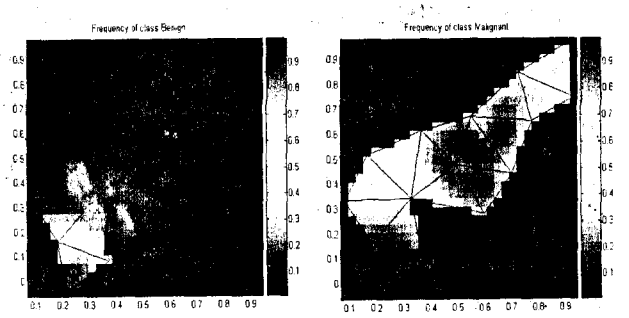

图 2　诊断结果比较图(左：良性；右：恶性)

为了便于更明显的观察，将使用概率分布图对参数进行显示，如图 3 所示。从图 3 可以看到明显的分类，在三维图形的立体交汇处有凹陷进去的分界线。通过图 2、图 3 可以直观对乳腺癌症细胞做出初步判断，当 9 个测试指标都处于 0.10 到 0.40 之间为良性，当 9 个测试指标都处于 0.40 到 1.00 之间为恶性。当然要精确地诊断出乳腺癌症细胞的性质，则需要对细胞的每个测试指标进行分析计算。由于篇幅所限，只针对 X1 、X2 与 X3 进行研究。如图 4 所示。

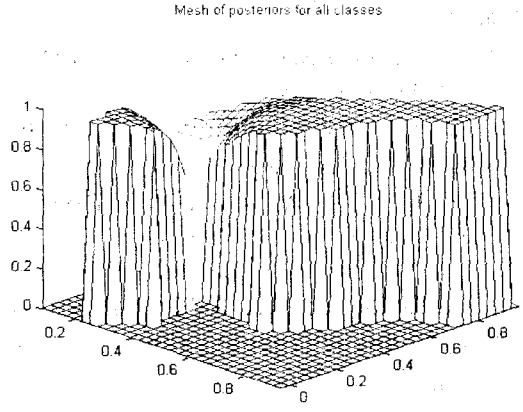

图 3　神经元的概率分布图

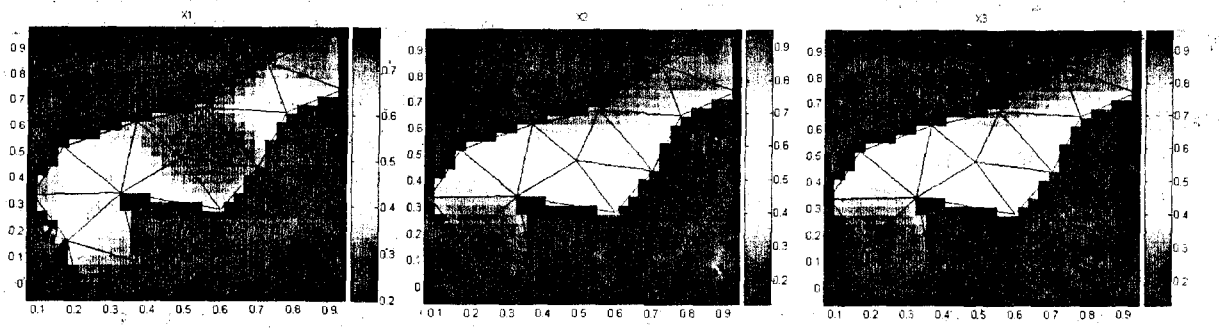

图 4　单个输入变量的特征图

从图中显示的结果看，此系列图表示的是 3 个输入变量(X1 - X3)在训练机制中的成分分析。分别显示了单个变量的维的分布情况以及值的大小分布。其中，网络下部靠近原点的深色区域代表神经元模式稀疏；网络上部深色区域则代表神经元模式稠密。而且从图中还可以判断出变量之间的相关性，比如 X2 与 X3 自相关。而且从图中的颜色分布可以看出，当 X1 的值保持在 0.5 到 0.7 或 0.8 到 1.0 之间时，诊断结果为恶性，其余情况为良性。根据此方法也可以对其余变量做出诊断分析。最后通过测试数据集诊断结果的有效性验证，进行预测判断的精度达到 91.333%。

3 结束语

通过大量测试参数的分析，可以看出 GCS 网络操作过程简便易行，只要将有关数据提供给网络，网络就会通过自身训练、学习，最后根据人的意愿输出相应的结果，根据这些结果并结合实际情况做出更准确的判断。但是 GCS 方法也具有一定的局限性，由于其采用非规范的动态图结构，GCS 执行起来比预定规则结构网络更为困难。另外在参数选取、样本数量等方面也要求更深入地进行研究。在测试过程中，应该注意收集更多更全面的病例数据。GCS 网络在医疗诊断中的各个病例的应用还有待于进一步的深入研究，在以后的科研工作中还要重点研究参数选取，以及拓扑的改进问题。

参考文献：

[1] 周国亮，刘希玉. 一种基于进化算法的 BP 神经网络优化方法及应用[J]. 计算机技术与发展，2007，17(8)：58 - 60.

[2] 杜华英，赵跃龙. 人工神经网络典型模型的比较研究[J]. 计算机技术与发展，2006，16(5)：97 - 99.

[3] 郭海强. 肺癌诊断模型的研究[J]. 中国卫生统计，1997，14(5)：238 - 245.

[4] 吴拥军，吴逸明，屈凌波，等. 人工神经网络在肺癌诊断中的应用研究[J]. 中国微生物和免疫学杂志，2003，23(8)：135 - 139.

[5] 陈新平. 临床医学中的神经网络技术[J]. 中国现代医学杂志，2003，13(9)：85 - 90.

[6] 樊晓干，彭 展，杨胜跃. 基于多层前馈型人工神经网络的抑郁症分类系统研究[J]. 计算机工程与应用，2004，40(13)：205 - 215.

[7] Cheng Guojian，Zell A. Externally Growing Cell Structures for Data Evaluation of Chemical Gas Sensors，Neural Computing，Applications[J]. Neural Computing & Applications，2001，10(1)：35 - 43.

[8] Fritzke B. Growing cell structures - a self - organizing network for unsupervised and supervised learning[J]. Neural Networks，1994，7：1441 - 1460.

(上接第 227 页)

[2] Carter S. The New Language of Business SOA&Web2.0[M]. 北京：清华大学出版社，2007.

[3] 彭 政，聂瑞华. 数字校园中基于 SOA 的数据交换平台设计[J]. 科技广场，2008(3)：72 - 74.

[4] 方蔚涛，杨 丹，李 珩，等. 数字化校园信息门户的设计研究[J]. 计算机科学，2007(3)：9 - 11.

[5] Luo Min，Goldshlager B，Zhang Liang - Jie. Designing and Implementing Enterprise Service Bus(ESB) and SOA Solutions [M]. Washington，DC，USA：IEEE Computer Society，2005.

[6] 刘迎春，兰雨晴，于乐乐. ESB 中的数据交换技术[J]. 计算机系统应用，2005，10：42 - 45.

[7] 许 鑫，苏新宁，吴乃刚. 高校共享数据中心平台的设计与实现[J]. 现代图书情报技术，2005(6)：48 - 53.

[8] 张泽文，聂瑞华，刘海星，等. 基于. NET 的协同工作模式研究与实现[J]. 华南师范大学学报：自然科学版，2007(1)：43 - 47.

(上接第 230 页)

改进建议：采用多代理技术(Multi - agent)。目前已经尝试利用多代理技术对系统进行改进，还在调试当中，今后将继续努力把这一工作完成。

参考文献：

[1] 姚志强，陈文博，沈 媛. 基于 Web Service 架构的 CSCW 应用研究[J]. 计算机工程与应用，2005(3)：132 - 136.

[2] Riboulet V，Marin P，Leon J. Towards a New Set of Tools for a Collaborative Design Environment[J]. Computer Supported Cooperative Work in Design，2002，9(25 - 27)：128 - 133.

[3] 孙卫琴，李洪成. Tomcat 与 Java Web 开发技术详解[M]. 北京：电子工业出版社，2005.

[4] 王恩涛，李 祥. 基于 Socket 的手机与数据库服务器通信的研究[J]. 计算机技术与发展，2007，17(2)：81 - 84.

[5] 刘邦桂，李正凡. 用 Java 实现流式 Socket[J]，华东交通大学学报，2007(5)：110 - 112.

[6] Ayers D，Bell J，Calvertbettis C. Professional Java Data[M]. [s.l.]：Publishing House of Electronics Industry，2002.

[7] 马喜春，张曾科. 基于 Socket 进行通用的网络通信程序设计[J]，实验技术与管理，2005(3)：58 - 61.

[8] 范云芝，陈树平. 利用 Socket 实现基于 Web 的远程监测系统[J]，陕西工学院学报，2005(1)：53 - 55.

[9] Microsoft Corporation. Developing Client/Server Applications with Visual Basic[M]. USA：Microsoft Press，1996.