·基础理论与方法·

人工神经网络应用于糖尿病/糖耐量异常的疾病分类研究

钱玲 施侣元 程茂金

【摘要】 目的 探讨人工神经网络(ANN)用于疾病分类研究的前景。方法 利用某矿区 1996 年糖尿病现况调查资料,采用学习向量量化(LVQ)网络和判别分析方法进行糖尿病/糖耐量(DM/IGT)异常/正常状态的判别比较;同时人为设置变量缺损值,检验 LVQ 网络对缺失数据的适应性。结果 LVQ 网络结构为 25→13→3;网络判断准确率为96.98%,对血糖异常者的正确判断率为92.45%。利用逐步判别分析建立的含 11 个变量的判别方程的判断准确率为87.34%,对血糖异常者的正确判断率为85.53%。LVQ 网络对带缺失项样本的误判比例为 1/30,判别分析则为 7/30。结论 利用 LVQ 网络进行疾病分类预测,不仅能获得更好的预测效果,而且对资料的类型、分布不作任何限制,也不需要对分析变量做任何处理,还能很好地处理带缺失项的资料,是一种很好的流行病学分类预测新方法。

【关键词】 人工神经网络;学习向量量化网络;糖尿病/糖耐量异常;疾病分类

Study on the application of artificial neural network on diabetes mellitus/insulin-glucose tolerance classification *QIAN Ling**, *SHI Lv-yuan*, *CHENG Mao-jin*. **National Health Education Institute*, *Chinese Center for Disease Control and Prevention*, *Beijing* 100011, China

【**Abstract**】 **Objective** To discuss the potential application of artificial neural network (ANN) on the epidemiological classification of disease. **Methods** Learning vector quantization neural network (LVQNN) and discriminate analysis were applied to data from epidemiological survey in a mine in 1996. **Results** The structure of LVQNN was 25→13→3. The total veracity rates was 96.98%, and 92.45% among the abnormal blood glucose individuals. Through stepwise discriminate analysis, the discriminate equations were established including 11 variables with a total veracity rate of 87.34%, but was 85.53% in the abnormal blood glucose individuals. Further analysis on 30 cases with missing values showed that the disagreement ratio of LVQ was 1/30, lower than that of discriminate analysis of 7/30. **Conclusions** Compared to the conventional statistics method, LVQ not only showed better prediction precision, but could treat data with missing values satisfactorily plus it had no limit to the type or distribution of relevant data, thus provided a new powerful method to epidemiologic prediction.

【**Key words**】 Artificial neural network; Learning vector quantization neural network; Diabetes mellitus/insulin glucose tolerance; Classification of disease

疾病分类研究是根据个体所具有的状态和特征,利用模式之间的相似程度直接识别疾病状态。使用的识别及分类方法很多,如 logistic 回归分析、主成分回归分析、判别分析、医学专家系统等,然而由于方法上的局限性,在使用过程中往往会遇到一些难以解决的问题或者步骤繁琐。人工神经网络(ANN)是为模仿人类的高级智能活动而设计的一种结构系统,因其具有自组织、自学习、自适应和并行处理(响应速度快)等特点以及很强的输入输出非线性映射能力和易于学习和训练的优点,被广泛应用于非线性系统辨识、模式识别、模式分类等方面[1,2],成为当前具有智能模式识别能力的工具之一。应用 ANN 于疾病的分类和诊断国内外已有报道[3-6],然而目前大多数是以临床资料为基础,预测患者的疾病状态,制定适当的诊治方案。本文以流行病学资料为基础,探讨利用 ANN 进行疾病分类研究的效果及特点。

基本原理

1. 基本结构:

作者单位:100011 北京,中国疾病预防控制中心健康教育所(钱玲);华中科技大学同济医学院公共卫生学院(施侣元、程茂金)

(1)人工神经元:ANN 作为模拟生物神经网络的一种信息处理系统,其基本处理单元即人工神经元(或称节点、单元、神经元)是人脑神经元的简化和模拟。人工神经元是一种多输入、单输出的非线性单元,数学模型为

$$y=f\left(\sum_{i=1}^{N}w_ix_i-\theta\right)$$

即神经元将接收信息 $\{x_i\}$ 与互连权重 $\{w_i\}$ 的点乘积求和构成其总输入,经函数 $f()$ 的作用,产生输出,函数 $f()$ 通常为非线性。

(2)人工神经网络:ANN 通常包含一个输入层(input layer)、一个或几个隐含层(hidden layer)和一个输出层(output layer),每层均由若干神经元构成。输入层神经元只从外界接受信息,不具备运算功能,也不受其他神经元的作用;输出层神经元以从其前一层接受的信息作为本身的输入,将整个网络对输入数据的处理结果经过某种数学转换后传给外界,是整个网络的信息出口;介于输入层和输出层之间的隐含层是整个网络的核心,其结构与神经网络的性能密切相关。ANN 的层与层之间是通过权重相互连接的,层内的各神经元不发生连接,每一层中各个神经元与权重相乘后在下一层中完成累加及函数转换后输出。

(3)学习向量量化(learning vector quantization, LVQ)网络:本研究采用的 LVQ 网络由输入层、竞争层(隐含层)和线性输出层三层组成。输入层和竞争层的神经元进行全互连连接,竞争层和线性输出层的神经元则是部分连接,每个线性输出层神经元则只与竞争层中对应的竞争获胜神经元连接。信息传递过程是由输入层到输出层的单向传递。

2. 工作原理:与人脑一样,ANN 是通过学习或训练来获得“知识”或“经验”从而解决预测或模式识别等问题的。ANN 的学习或训练即是按照一定的规则不断调整权重大小以最大程度地进行样本学习和识别、完善网络性能的过程,其所有“经验”均以权重形式保存。ANN 的学习方式包括有教师(监督)学习和无教师(无监督)学习两种,前者除向网络提供输入信息以外,还提供理想输出(即实际结果),网络将根据实际输出与理想输出之间的差值调整网络系数;而后者只向网络提供输入信息,由网络进行自组织学习和识别。

LVQ 网络是自组织特征映射模型的一种改进,其采用的 LVQ 是在监督状态下对竞争层进行训练的一种学习规则。所谓竞争层训练指的是竞争层的各个神经元在接收到输入样本信息后相互之间进行竞争,最终只有一个或几个神经元活跃,以适应当前的输入样本,这些竞争胜利的神经元就代表了当前的输入样本的分类模式,并修改与其相联的连接权重。在 LVQ 学习规则中,竞争层神经元依赖输入向量之间的距离,自动学习把输入向量进行分类,并将结果输出到线性层神经元上,由线性层组合传递到用户定义的目标分类上,从而将输入向量中与目标向量相近的分离出来[7]。

3. 网络结构及参数的确定:

(1)网络结构的选择:关于 LVQ 网络输入变量的选择,目前尚缺乏与 ANN 配套的较成熟的变量选择方法。有人提出利用逐步判别分析进行变量筛选[8],以便于与传统判别分析比较;也有人提出采用 logistic 回归分析筛选出具有显著意义的单因素变量。后者则基于“logistic 回归模型在结构上等同于以 logistic 函数作为激发函数的单层前馈网”的原理而得到经常应用[9]。以 logistic 回归分析筛选出的变量个数作为网络输入层神经元的个数;输出神经元个数则等于期望输出的分类数。作为 LVQ 网络核心的竞争层神经元数目的确定目前尚无公认的有效方法。本研究采用试法根据交叉证实组 ROC 曲线下面积来确定合适的网络结构[10],当特异度为 100%,总准确率在 95%以上时,网络预测标准误差(standard error of prediction, SEP)最小[11],为 10%。

(2)网络的初始化及学习训练:网络的初始化即是确定网络各连接权的初始权值和偏差。LVQ 网络根据输入样本、竞争层神经元个数以及目标分类来进行网络初始化,得到输入层神经元和竞争层神经元以及竞争层神经元和线性输出层神经元的初始连接权值。预先设立的学习速率规定了在网络学习训练中权值调整的幅度,LVQ 网络将根据初始权值、输入样本、竞争层实际输出、竞争层的目标输出以预定的学习速率进行网络学习和训练,得到调整的权值。训练终点的确定则是利用交叉证实组和 SEP,以防止网络的过度训练和保证网络良好的预测性能。

实例分析

为了进一步探讨及评价在流行病学调查资料基础上,ANN 用于疾病分类研究的效果和特点,我们以某矿区 1996 年糖尿病现况调查资料为基础,采用

LVQ 网络进行疾病分类研究，并与当前医学领域中广泛应用的判别分析相比较。

1. 研究对象分组：

将矿区调查中筛选出的 177 例糖尿病(DM)患者、141 例糖耐量低减(IGT)者及3 062名血糖正常者(NGT)随机均分成两组，第一组人数为1 689人，第二组为1 691人。

第一组中按 33% 的抽样比随机抽取 30 例 DM 患者、20 例 IGT 者及 507 例 NGT 者共 557 例构成交叉证实组，剩下的1 132例研究对象为训练组，两组共同用于选择 ANN 网络结构及判断训练终止点；第二组的1 691名研究对象为测试组，用于评价网络的拟合效果；为了检验神经网络对缺失数据的适用性，我们在两组中随机选择了 30 例样本，按 logistic 回归分析中获得的因子顺位顺序将样本中的元素逐一地设置为缺损值，用已训练好的神经网络作分类预测。

2. ANN 分析结果：

(1)变量筛选结果：利用某矿区的全部现况调查资料，以个体状态(DM、IGT、NGT)作为因变量进行单因素非条件 logistic 回归分析，在0.05水平上筛选出有统计学意义的单因素变量(表 1)。

表1 糖尿病/糖耐量异常单因素非条件 logistic 回归分析结果

变 量	β值	*OR* 值	*P* 值
腰臀比	9.797 1	999.000	<0.000 1
视网膜病史	2.225 8	9.261	<0.000 1
离退休者	2.072 0	7.941	<0.000 1
高脂血症病史	1.787 6	5.975	<0.000 1
高血压病史	1.650 9	5.212	<0.000 1
脑血管疾病病史	1.642 3	5.167	<0.000 1
冠心病史	1.402 0	4.063	<0.000 1
肾病史	1.412 0	4.104	<0.000 1
肝病史	1.009 0	2.743	<0.000 1
其他患病史	0.967 0	2.630	<0.000 1
DM 家族史	0.907 6	2.478	<0.000 1
服务人员	0.738 1	2.092	0.032 6
家庭妇女	0.748 5	2.114	0.000 1
行政干部	0.478 8	1.614	0.002 5
体质指数	0.190 7	1.210	<0.000 1
年龄	0.090 9	1.095	<0.000 1
脉率	0.070 8	1.073	<0.000 1
舒张压	0.037 9	1.039	<0.000 1
收缩压	0.031 2	1.032	<0.000 1
文化程度	−0.287 8	0.750	0.000 1
饮酒史	−0.3174	0.728	0.003 6
职业体力活动	−0.381 1	0.683	<0.000 1
吸烟史	−0.382 5	0.682	0.000 3
性别	−0.762 1	0.467	<0.000 1
工人	−1.402 4	0.246	<0.000 1

(2)LVQ 网络结构选择结果：以单变量筛选出的 25 个变量为网络输入变量，输出变量为 DM/IGT/NGT 三种状态。根据交叉证实组 ROC 曲线下面积和 SEP 选择网络结构，LVQ 网络输入层、竞争层、线性层神经元的个数分别为 25、13、3。

(3)LVQ 网络训练及拟合结果：LVQ 网络训练的学习速率为0.1；到达训练终点时，训练总步数为 15 690步，网络预测结果见表 2。

表2 LVQ 网络分析结果

实际类别	网络输出分类			合计
	DM	IGT	NGT	
DM	62	24	2	88
IGT	15	46	10	71
NGT	0	0	1 532	1 532
合计	77	70	1 544	1 691

3. 判别分析结果：对单因素分析有意义的变量，经逐步判别分析，共筛选出 11 个有显著意义的判别指标，分别是腰臀比、DM 家族史、肾病史、高脂血症病史、高血压病史、脉率、职业性体力活动、职业、舒张压、体重指数、年龄。以此建立判别方程对测试样本的判别结果见表 3。

表3 判别分析结果

实际类别	判别分类			合计
	DM 阳性	IGT 阳性	阴性	
DM	59	22	7	88
IGT	12	43	16	71
NGT	95	62	1 375	1 532
合计	166	127	1 398	1 691

4. 缺失项对预测效果的影响分析：30 例带有缺失项的数据比较分析结果，以及利用 SEP 判断各变量对网络分类预测影响的结果分别见表 4、图 1。

表4 带缺失项样本的分析结果

样本序号	网络分析	判别分析	样本序号	网络分析	判别分析
1	F	F	16	T	T
2	T	F	17	T	T
3	T	F	18	T	T
4	T	F	19	T	T
5	T	F	20	T	T
6	T	F	21	T	T
7	T	F	22	T	T
8	T	T	23	T	T
9	T	T	24	T	T
10	T	T	25	T	T
11	T	T	26	T	T
12	T	T	27	T	T
13	T	T	28	T	T
14	T	T	29	T	T
15	T	T	30	T	T

注：F 表示带缺失项样本在分析中的结果与实际分类不一致；T 表示带缺失项样本在分析中的结果与实际分类一致

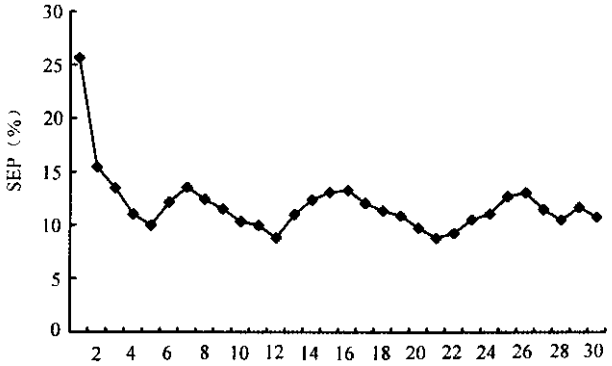

图1 带缺失项的独立样本预测结果

讨 论

判别分析是用于对具有明确分类的疾病状态进行预测的最常用方法，包括二类判别、多类判别、逐步判别、典则判别及非参数判别分析等，每一种分析方法均有其各自的使用条件。理论上，如果原资料或经过转换后的资料符合参数判别分析的适用条件，即能据此得出合理的结论；但是实际应用中所获取的资料并不总是已知其分布类型或者说并不是总呈正态或近似正态分布的，而往往是多种分布类型并存，此时参数判别分析有其使用局限，应用非参数统计方法又会导致资料的利用率较低，统计效率降低[12]。

对 DM/IGT 而言，影响疾病状态的因素众多，作用方式复杂，以流行病学资料为基础利用传统线性判别函数这种“刚性”方法进行疾病状态预测，就存在很大的局限性。以非线性大规模并行分布处理为特点的 ANN 理论则突破了传统的线性处理模式，避开了复杂的参数估计过程去解决一系列变量关系不能精确地用函数表达的分类与回归问题，因而成为人们探索和研究某些复杂大系统的有力工具[3,6]。其中 LVQ 又因其网络竞争层能根据给定的输入、输出特征自动学习适应样本特点，并依据输入变量之间的距离进行分类，同时分离出和输出向量相近的输入向量[13]，而广泛应用于模式分类预测。LVQ 网络应用在疾病分类预测时，对资料的类型、分布不作任何限制，也不需要对分析变量做任何处理，同时对研究者的专业知识不作太高要求，是一种既方便又可充分利用信息的方法[14]。本研究的 LVQ 网络拟合结果显示：ANN 可获得比传统的判别分析法更好的预测效果。对于同一测试样本，其判断准确率和识别血糖异常者的正确率均高于判别分析，分别为96.98%[(62+46+1 532)/1 691]和92.45%[(62+24+15+46)/(88+71)]，高于判别分析的87.34%和85.53%。

须明确的是，ANN 是根据经验进行判断、解决问题的，网络训练样本中不同类别样本的数目可直接影响到网络对该类样本的识别率。本研究所用的训练样本中正常个体占绝大部分90.37%(1 023/1 132)，因此在对测试样本的检验中该网络对正常个体的识别准确率为 100%(1 532/1 532)，而对异常个体的识别能力则相对较低，也就是说网络对血糖正常者的识别能力及判断准确率均比较稳定；而对血糖异常者的识别与待判资料的内部构成有关。尽管如此，就 ANN 本身来讲，其预测能力是无限精确的。在推广使用中可结合考虑不同疾病的严重性、可否得到有效治疗以及经济可行性等因素来调整训练样本组成，建立不同灵敏度及特异度的判别模型，满足不同的需要。

此外，在流行病实际调查中，由于调查的规模、对象的依从等原因，常常会出现样本数据不完整的情况。传统的统计模型通常是将其删去，以防止对总体预测的影响，降低预测能力，这样做无疑会损失一部分信息。然而，ANN 由于有较强的容错性以及擅长处理模糊、不完全和不精确数据的能力，能够很好的处理带有缺失或干扰项目的资料[15,16]。本研究对 30 例带有缺失项的数据比较分析可见缺项对 ANN 网络分析结果几乎没有影响，不一致的比例仅为1/30，除了第一例缺失重要预测变量腰臀比的样本之外，其他均得到一致结论；相对而言判别分析则要逊色得多，在前 7 个缺失样本中全部误判，缺失项目对预测结果影响较大。

ANN 的影响参数较多，通常需通过实验来优化和选择网络参数。本文采用预测 SEP 考察网络参数的影响，并根据样本 SEP 对网络正常 SEP 的偏差程度来寻找影响网络分类判别的主要因素。若某变量对分类的影响较小，则将该变量含量置为缺项时，计算结果的 SEP 偏离正常 SEP 值较小；反之，若计算结果的 SEP 偏离正常 SEP 值愈大，则表明该变量对分类的影响愈大。本研究所建立的网络对大部分带缺失项样本都能正确判断出来，样本 SEP 偏离正常 SEP(10%)较小；只有第一例缺失腰臀比的样本 SEP 偏离正常 SEP 值较大，说明腰臀比对正常人与血糖异常者的分类起着关键性的作用。如果未知样本中缺少了此类变量，则不能做出正确的分类。利用 ANN 对带缺失项样本的预测结果显示了 ANN

除了可以根据主要影响因素对血糖异常者进行识别外，还可以考察各影响因素对分类判别的影响。对该分类判别影响较小的因素，当有缺损值时，仍可做出正确的判别[11]。

本文采用 LVQ 网络进行疾病分类研究，并与传统的判别分析做比较，发现 LVQ 网络不仅能获得更好的预测效果，而且对被研究资料无任何要求，在应用时不需考虑资料的特征或进行资料转换，非常直接方便。LVQ 网络同样能很好的处理带缺失项的资料，是一种很好的流行病学分类预测新方法。然而尽管 ANN 具有常规方法所不可替代的优势，但 ANN 模型建立较为繁琐，训练时间要求较长，尤其是目前 ANN 方法在流行病学中的应用尚处于起步阶段，还存在一些问题如网络参数以及网络结构的选择方法、如何防止训练过度等尚需要更深入探讨。

参 考 文 献

1 Chen S, Billings SA. Neural networks for non linear dynamic system modeling and identification. Int J Control, 1992, 56: 291-319.

2 谭宇明，苏开才，毛宗源. 基于神经网络的带补偿作用的机器人逆动力学控制. 控制理论与应用，1997，14：7-11.

3 王小如，朱尔一. 分析化学进展. 南京：南京大学出版社，1994. 925.

4 蔡煜东，朱建中，甘骏人，等. 人工神经网络在冠心病患者血液分析中的应用. 分析化学，1992，20：885.

5 Mobley BA, Schechter E, Moore WE, et al. Predictions of coronary artery stenosis by artificial neural network. Artif Intell Med, 2000, 18: 187-203.

6 Sonke GS, Heskes T, Verbeek AL, et al. Prediction of bladder outlet obstruction in men with lower urinary tract symptoms using artificial neural networks. J Urol, 2000, 163: 300-305.

7 施洋，李俊. MATLAB 语言工具箱——TOOLBOX 实用指南. 西安：西北工业大学出版社，1999.

8 邓小元，李坤成，刘树良. 人工神经网络在 Alzheimer 病的 MRI 诊断研究中的应用. 中华放射学杂志，1998，32：812-816.

9 Duh MS, Walker AM, Ayanian JZ. Epidemiologic interpretation of artificial neural network. Am J Epidemiol, 1998, 147: 1112-1122.

10 Duh MS, Walker AM, Pagano M, et al. Prediction and cross-validation of neural networks versus logistic regression: using hepatic disorders as an example. Am J Epidemiol, 1998, 147: 574-579.

11 Zhang ZY, Liu SD, Ding BJ, et al. Artificial neural network applied to diagnosis of lung cancer. Chem J Ch Univ, 1998, 19: 530-533.

12 陈心广，尹平. 医学科研设计与数据分析. 武汉：武汉大学出版社，1997.

13 Kohonen T. The self-organizing map. Proc IEEE, 1990, 78: 1464-1480.

14 Wei JT, Zhang Z, Barnhill SD, et al. Understanding artificial neural and exploring their potential applications for the practicing urologist. Urology, 1998, 52: 161-172.

15 Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regresson for predicting medical outcomes. J Clin Epidemiol, 1996, 49: 1225-1231.

16 Braitman LE, Davidoff F. Predicting clinical states in individual patients. ANN Intern Med, 1996, 125: 406-412.

（收稿日期：2003-01-16）
（本文编辑：张林东）

·消息·

《中华流行病学杂志》2004 年征订启事
月刊·全铜版纸印刷·定价不变·个人订阅优惠

《中华流行病学杂志》是由中华医学会主办的流行病学及其相关学科的高级专业学术期刊、国内预防医学和基础医学核心期刊、国家科技部中国科技论文统计源期刊，并被美国国立图书馆医学文献联机数据库收录。读者对象为预防医学、临床医学、基础医学及流行病学科研与教学工作者。征稿内容：重点或新发传染病现场调查与控制；慢性非传染病的病因学及流行病学调查（含社区人群调查）、干预与评价；环境污染与健康；食品安全与食源性疾病；流动人口与疾病；行为心理障碍与疾病；分子流行病学、基因学与疾病控制；我国西部地区重点疾病的调查与控制等。本刊设有述评、重点原著、疫情监测、现场调查、实验研究、临床流行病学、疾病控制、基础理论与方法、国家重点课题总结、文献综述、问题与探讨等重点栏目。

本刊每期 84 页，全年出版 12 期，每期定价 9 元（含邮费），全年 108 元，由全国各地邮局统一订阅，邮发代号：2-73。本刊编辑部常年办理邮购。2004 年本刊执行个人直接订阅优惠办法（2004 年全年 12 期优惠价 90 元，优惠订阅截止日期 2004 年 1 月 31 日，此方法仅限直接向编辑部邮购的个人）。地址：北京昌平流字五号《中华流行病学杂志》编辑部，邮编：102206，电话（传真）：010-61739449，E-mail：xbonly@public3.bta.net.cn 欢迎广大读者踊跃投稿，积极订阅。

本刊编辑部