·基础理论与方法·

# 神经网络在预测糖尿病发病中的应用研究

高蔚 施侣元 董福霞

用机器模拟人类智能活动是人类长期以来梦寐以求的理想，人工神经网络(artificial neural network, ANN)，简称神经网络(neural network, NN)正是这样一种模型：它巧妙地将生物神经网络(biological neural network, BNN)的结构和工作原理用数学形式模拟出来，使其具有人脑的某些功能。该方法自上个世纪 80 年代重新兴起以来，已在诸多领域获得成功应用[1]。在流行病学中也逐渐受到重视，关于疾病预测方面的研究虽有一些，但多是以临床资料为基础[2-4]，基于流行病学资料的疾病预测到目前为止还未见报道，我们将以某地糖尿病流行病学调查资料为基础对这一方法在流行病学个体疾病发病预测中的应用潜力及特点做一简要介绍。

## 一、方法及原理

1. 基本结构：与 BNN 类似，NN 的基本组成单位是神经元(neuron，又称结点 nodes)，对应于 BNN 的神经细胞，网络信息的加工、处理都是在其中完成的。它的三个基本组成要素是为连接权、求和单元以及非线性激发函数(activation function, transfer function)，分别模仿生物神经网络的突触连接特性、神经细胞的整合特性以及阈值特性。其结构示意图见图 1。由若干个神经元通过权值按照一定的方式连接起来就组成了不同类型的 NN[5]。本文所采用的是一种在模式识别和模式分类中经常用到的 LVQ 网，其结构示意图见图 2[6]。这属于结构相对简单的一种 NN，由输入层、竞争(隐含)层以及线性(输出)层共三层构成，信息流向为单向，其中输入层和竞争层神经元为全连接，而竞争层与线性层神经元则部分连接，每个线性层神经元与竞争层神经元的不同组相连接。

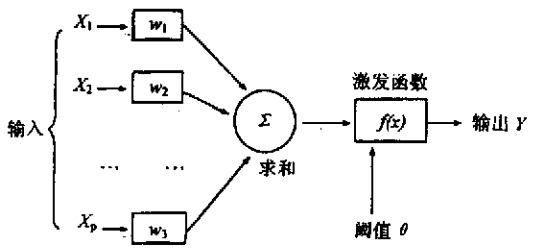
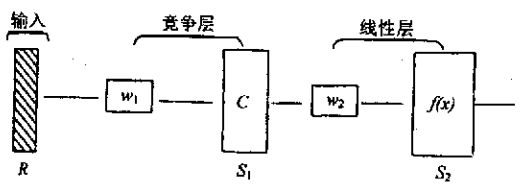



图1 神经元模型




$R$ 为输入数 $S_1$ 为竞争层神经元数 $S_2$ 为线性层神经元数

图2 LVQ 网络结构图

2. 工作原理：与 BNN 一样，NN 是以“知识”(或“经验”)为基础解决问题，而且它的“知识”也是通过学习获得，并以权值的形式贮存在网络之中的，NN 的学习即网络根据一定的规则调整内部权值的行为，学习算法则是实现权值调整的具体方法。学习方式主要分为监督学习和无监督学习两种，其中监督学习除需向网络输入外，还需提供期望输出(正确答案)，NN 根据期望输出与实际输出之间的差值(误差信号)来调节系统参数，而无监督学习则只向网络输入，不提供期望输出。不断重复的学习过程即训练，训练是使网络获得良好预测性能的必经过程。

LVQ 网是对自组织特征映射(self-organizing feature map, SOFM)的一种改进模式，其采用的 LVQ 算法可在监督状态下对竞争层进行训练，在竞争层中网络将根据给定的输入、输出特征自动学习适应样本特性并对输入样本分类，其具体工作步骤为：①每输入一个样本 $x$ 后，在输出阵列中找出具有最大输出的单元 $c$；②设 $x$ 所属类别已知为 $r$，而在第一阶段学习中 $c$ 的类别为 $s$，则权值按如下公式调整：

$$\begin{cases} w_c(t+1) = w_c(t) + \mu(t)[x(t) - w_c(t)] & r = s \\ w_c(t+1) = w_c(t) - \mu(t)[x(t) - w_c(t)] & r \neq s \\ w_i(t+1) = w_i(t) & i \neq c \end{cases}$$

由上式可见，若对 $x$ 分类正确，则使 $c$ 的权更靠近 $x$，否则远离。对不是最大输出的单元，权值不动[6]。

NN 的三大特点为高度并行性、高度的非线性全局作用以及良好的容错特性，这使得 NN 不仅对流行病学问题有广泛的适用性，而且对于大规模流行病学调查中易出现的缺失数据以及部分错误信息也有很一定的处理能力[5]。

3. 网络结构的选取、训练及性能评价：由于目前尚缺乏与 NN 分析相配套、较成熟的变量选择方法，而在结构上，logistic 回归模型又等同于以 logistic 函数作为激发函数的单层前馈网[7]，因此为减少运算量即缩短运算时间，先采用 logistic 回归分析筛选出具有显著性意义的单因素变量，将这些变量作为 NN 的输入用于分析，输入层神经元数目与筛出

作者单位：430030 武汉，华中科技大学同济医学院流行病学教研室(高蔚、施侣元) 河南省平顶山市矿区中心医院(董福霞)

第一作者现在单位：510632 广州，暨南大学医学院(E-mail：gwps@ynmail.com)

的单因素变量个数等同，输出神经元数目与期望的目标输出分类一致；LVQ 网络结构的设计中竞争层是直接关系到将来网络性能好坏的核心因素，目前还缺乏公认有效的方法，本研究结合前人研究成果和实验测试结果确定竞争层神经元规模。

LVQ 网各连接权的初始权值是随机分配的，同时需设定学习率以确定权值调整的幅度。与其他 NN 一样，LVQ 网也存在过度训练问题，即网络拟合性能在达到某个临界点后如果继续训练虽然模型对训练样本的拟合能力仍会继续提升，但却使模型对测试样本的预测性能下降。交叉证实法（cross-validation）是目前认为较好的方法之一[8,9]，其具体操作过程为：将训练组随机分为若干子集，随机取其中一个子集作为交叉证实组，其他剩下各组用于训练网络，在训练过程中每输入一定数量的训练样本用证实组经常检查网络的拟合性能，如此反复并用随机的方式选择不同的子集作为证实组，当网络对证实组的总预测性能开始下降或达到预定目标时即停止训练。训练结束后，还需用一组在网络训练期没有“见”过的样本集即测试组对网络性能作出评价，若预测性能符合预定目标即可尝试实际应用。

## 二、实例分析

为进一步探讨及评价 NN 用于流行病学个体疾病预测的效果及特点，我们以3 222例（其中糖尿病患者 172 例，正常者 3 050人）人群糖尿病流行病学筛查资料为基础，采用 LVQ 网对糖尿病患病状态进行预测，并将其与当前医学领域中广泛应用的个体疾病状态预测方法——判别分析相比较。

1.对象分组：将糖尿病患者及正常者随机等分为两组，其中一组为训练组用以训练网络，另一组为测试组用于评价网络性能。训练采用交叉证实方式，综合前人研究成果[10]结合本文研究目标确定交叉证实组中患者及正常人的具体构成为 16/84。为判断缺失数据对预测效果的影响，在测试组中人为制造 20 例带有一个缺失变量的样本，缺失变量按 logistic 回归分析中获得的因子顺位顺序选取，即第一例缺失 β 值最大的变量项，第二例缺失 β 值第二大的变量项，依此类推。

2.NN 分析结果：

（1）变量筛选结果：经 logistic 回归分析在 0.05 水平上共筛选出 22 个有显著性意义的变量，见表 1。

（2）网络结构选择结果：输入层、竞争层神经元个数分别为 22、12，线性层神经元数为 2。

（3）网络训练及预测结果：网络训练学习率定为0.1，当对交叉证实组真阴性率为 100%，总准确率 > 95% 时停止训练。网络预测结果见表 2。

3.判别分析结果：经逐步判别分析，共筛选出 8 个有显著意义的判别指标，它们分别是腰臀比、职业、糖尿病家族史、脉搏、肾病史、高血脂症史、高血压史、年龄。将这 8 个指标用于建立判别方程，结果见表 3。

4.缺失项对预测效果的影响分析：由表 4 可见缺失项对网络的分析结果几乎没有影响，不一致的比例仅为 1/20，除了第一例缺失重要预测变量腰臀比的样本之外，其他均在两种状态下获得一致结论，而判别分析则在前 6 个缺失样本中全部误判，缺失项目对预测结果影响较大。

表1 单因素 logistic 回归分析结果

| 变量 | β值 | *OR* 值 | *P* 值 |
|---|---|---|---|
| 腰臀比 | 1.945 | 6.995 | 0.000 1 |
| 高血脂症史 | 1.791 | 5.994 | 0.000 1 |
| 高血压史 | 1.690 | 5.421 | 0.000 1 |
| 脑血管病史 | 1.507 | 4.513 | 0.000 1 |
| 收缩压 | 1.239 | 3.452 | 0.000 6 |
| 体重指数 | 1.128 | 3.089 | 0.000 4 |
| 肝病史 | 1.003 | 2.727 | 0.017 1 |
| 吸烟史 | 0.668 | 1.950 | 0.000 4 |
| 文化程度 | 0.335 | 1.398 | 0.000 8 |
| 饮酒史 | -0.666 | 0.514 | 0.007 2 |
| 经济收入 | -0.263 | 0.769 | 0.031 3 |
| 脉　搏 | 1.927 | 6.866 | 0.000 1 |
| 肾病史 | 1.739 | 5.690 | 0.000 1 |
| 糖尿病家族史 | 1.688 | 5.407 | 0.000 1 |
| 冠心病史 | 1.445 | 4.242 | 0.000 4 |
| 其他病史 | 1.202 | 3.328 | 0.000 9 |
| 舒张压 | 1.061 | 2.889 | 0.000 2 |
| 年　龄 | 0.786 | 2.195 | 0.000 1 |
| 居住年限 | 0.358 | 1.430 | 0.000 2 |
| 性　别 | -0.948 | 0.388 | 0.000 1 |
| 职　业 | -0.324 | 0.723 | 0.000 1 |
| 职业性体力活动 | -0.190 | 0.827 | 0.000 4 |

表2 LVQ 网络分析结果

| 实际结果 | 网络预测结果 | | 合计 |
|---|---|---|---|
| | 正常数 | 异常数 | |
| 正常数 | 1 525 | 0 | 1 525 |
| 异常数 | 25 | 61 | 86 |
| 合计 | 1 550 | 61 | 1 611 |

表3 判别分析结果

| 实际结果 | 网络预测结果 | | 合计 |
|---|---|---|---|
| | 正常数 | 异常数 | |
| 正常数 | 1 251 | 274 | 1 525 |
| 异常数 | 28 | 58 | 86 |
| 合计 | 1 279 | 332 | 1 611 |

表4 带缺失值样本的分析结果

| 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 | D* | D | 11 | A | A |
| 2 | A** | D | 12 | A | A |
| 3 | A | D | 13 | A | A |
| 4 | A | D | 14 | A | A |
| 5 | A | D | 15 | A | A |
| 6 | A | D | 16 | A | A |
| 7 | A | A | 17 | A | A |
| 8 | A | A | 18 | A | A |
| 9 | A | A | 19 | A | A |
| 10 | A | A | 20 | A | A |

* 表示带与不带缺失项的样本在两次分析中结果不一致；

** 表示带与不带缺失项的样本在两次分析中结果一致

## 三、讨论

对具有明确分类的疾病状态进行预测最常用的是判别分析，它分为很多亚类[11,12]，每一种都有其各自的适用条件，应用时首先需对欲分析资料进行分析，根据资料特点选用相应的方法，而非参数方法虽然对变量的分布类型无特别要求，但其统计效率往往很低。这一过程还是建立在已知该问题采用判别分析的基础上，如果连这也不清楚，要从众多的统计学方法中选择合适的一种是非常不容易的，需要有很强的统计学背景，而方法选择过程又是必不可少的，因为如果对资料判断不准或对欲分析资料不甚了解造成误用，往往会导致错误结论。

LVQ 算法可在竞争层中根据给定的输入、输出特征自动学习适应样本特性并对输入样本分类[6,13]，这一特性使 NN 在进行疾病个体状态预测中，对研究者的专业要求有可能降低到最低限度，研究者只需知道如何解释结果，而不必担心由于资料特征或统计学方法的适用条件认识不足而犯原则性错误，它甚至可以不对分析变量做任何处理如将计量资料变成等级资料（这在许多预测技术中是必需的），是一个拿来就能用的方法，应用起来既方便又可充分利用样本中所含的信息[14]。

NN 属于非线性模拟系统，其特有的设计及工作原理使其对被分析的资料特性几乎没有任何限制，并行处理信息的方式又使它对残缺资料以及干扰（错误）信号具有一定的处理能力，这使其在流行病学应用中具有其他现有方法无法比拟的优势。本研究中 NN 预测与判别分析预测糖尿病的总错误率分别为1.6%（0 + 25/1 611）和18.7%（274 + 28/1 611），同时 NN 对正常个体的识别率达到了 100%，NN 的预测效果明显好于传统统计学方法。这一方面与 NN 的超强非线性拟合能力有关，另一方面也与 NN 的容错特性是密不可分的[5]。

NN 是根据‘经验’进行判断、解决问题的，在训练过程中如果它‘见’到的多为正常个体，则它识别正常个体的能力也会相应增强[15]，如本研究所用的训练样本中正常个体占绝大部分94.66%（1 525/1 611），因此在外推证实过程中该网络对正常个体的识别准确率达到 100%（1 525/1 525），对异常个体的识别能力则相对较差，需强调的是：从网络本身来讲，它的预测能力可以无限精确，对一组给定的样本来说，网络的最终预测结果与训练样本特性密切相关，这样就可通过调整训练样本组成以满足不同的需要。

此外，由于调查的规模较大以及调查对象的不配合等因素，流行病学资料中常常会有一些带有缺失项目的样本，对于这种资料，为防止这些样本对总体预测的干扰，降低预测效力，传统的模型往往是将其删去，但这样做无疑将损失一部分信息，这种情况在传统的预测模型中一直都没有更好的解决办法。由于 NN 的一个主要特点为对信息的处理及推理过程具有高度并行的特点，这一特性使网络中的每一个神经元都作为相对独立的功能单位对输出产生作用，但作用又很微小，这样即使输入信息中含有少量空缺项或干扰项，也不至于会影响整个网络的输出，因此能很好地处理带有缺失或干扰项目的资料，本研究的结果证明了这一点，这是该方法的另一个独特优点[16]。

尽管 NN 方法具有诸多优点，但 NN 方法还处于起步阶段，仍存在一些缺陷，如与常规方法相比，模型的建立及训练时间较长，同时许多关键问题如模型结构的选取、训练方式以及过度训练的防止等问题还需进一步探讨。随着对神经网络研究的不断深入，必将会获得更加广泛的应用。

## 参 考 文 献


1 Tewari A. Artificial intelligence and neural net-works: concept, applications and future in urology. Bri J Uro, 1997, 80(suppl 3): 53-58.

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

3 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.

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

5 李士勇. 模糊控制·神经控制和智能控制论. 哈尔滨：哈尔滨工业大学出版社, 1996. 77-102.

6 楼顺天, 施阳. 基于 Matlab 的系统分析与设计——神经网络. 西安：西安电子科技大学出版社, 1998. 20-22.

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

8 Diamantidis NA, Karlis D, Giakoumakis EA. Unsupervised stratification of cross-validation for accuracy estimation. Artificial Intelligence, 2000, 116: 1-16.

9 Koufakou A, Georgiopoulos M, Anagnostopoulos G, et al. Cross-validation in Fuzzy ARTMAP for large databases. Neural Networks, 2001, 14: 1279-1291.

10 Prechelt L. Automatic early stopping using cross validation: quantifying the criteria. Neural Networks, 1998, 11: 761-767.

11 苏炳华, 主编. 医学统计学及其软件包. 上海：上海医科大学出版社, 1996. 250-251.

12 曹素华, 主编. 实用医学多因素统计方法. 上海：上海医科大学出版社, 1998. 125-156.

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

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

15 Cross SS, Harrison RF, Kennedy RL. Introduction to neural net-works. Lancet, 1995, 346: 1075-1079.

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


（收稿日期：2001-11-21）
（本文编辑：张林东）