维普资讯 http://www.cqvip.com

中国卫生统计 2005 年 6 月第 22 卷第 3 期

BP 神经网络与 logistic 回归的比较研究

李丽霞[1] 王 彤[2] 范逢曦[2]

【提 要】 目的 通过与 logistic 回归分析的比较，探讨 BP 神经网络在判别分析中的应用。**方法** 设计合适的 BP 神经网络参数，采用 Levenberg-Marquardt 优化算法来避免 BP 算法收敛速度慢的缺点，采用了"早停止"(early-stopping)策略避免过度拟合(over-fitting)，并把 BP 神经网络和 logistic 回归的结果作比较。**结果** BP 神经网络在回代和前瞻性考核中都取得较好的结果，两者 ROC 曲线的比较也说明了这一点。**结论** BP 神经网络方法值得在医学研究，特别是判别分析、生存分析领域进一步应用并推广。

【关键词】 BP 神经网络 过度拟合 BP 算法 ROC 曲线 logistic 回归

BP 神经网络模型在应用中自变量可以是连续的，也可以是离散的，不需要考虑自变量是否满足正态性及变量间独立等条件，可以识别变量间复杂的非线性关系，尤其是用现有统计方法无法达到目的或效果不好时，采用此模型往往收到很好的效果。在统计学领域，主要将它应用在预测、判别分类问题中[1~3]。

方法介绍

BP 网是单向传播的多层前馈网络，采用典型的有师学习方式，其结构一般分为三层：输入层、输出层、隐含层(可有多个隐含层)，每一层可包含一个或多个神经元，其中每一层的每个神经元和前一层相连接，同一层之间没有连接。

BP 网的每一连线连接着两个神经元，并附有一个权值 W，权值表明了上一层神经元对下一层神经元的影响或连接强度，每一神经元的计算输出又是下一层所有神经元的输入，再用于计算它们的输出。BP 网所采用的传递函数一般是 sigmoid 型函数，例如：$f(x)=1/[1+\exp(-x)]$，BP 网可看成是一从输入到输出的高度非线性映射。BP 网络的学习过程包括正向传播过程(the forward phase)和反向传播过程(the backward phase)两部分。当给定网络的一个输入模式 X 时，它由输入层单元传到隐层单元，经隐层单元逐层处理后再送到输出层单元，由输出层单元处理后，产生一个输出模式 O，称为前(正)向传播，如果输出响应与期望输出模式有误差，不满足要求，那么就将误差信号沿原来的连接通路从输出层到输入层逐层传递，并修正各层的连接权值，直到误差信号最小，该过程称为反向传播。BP 算法(Back-propagation algorithm)为网络的学习提供了有力的方法，具体理论参见文献[1~2]。

实例分析

胆汁性肝硬化(PBC)是一种预后很差的病，判断病人能否存活 5 年在临床上有现实意义。本例通过建立 BP 网模型，来预测某一病人能否存活 5 年以上，并和传统的 logistic 回归的预测结果作比较。共收集 242 例胆汁性肝硬化病人，病人存活时间和许多指标有关，用 logistic 回归在 $\alpha=0.05$ 水平上逐步筛选变量(采用向前、向后法)，均选出 3 项有统计学意义的因素：年龄 X_1(岁)、胆红素 X_2(mg/dl)、白蛋白 X_3(mg/dl)，结果见表 1。

表 1 logistic 回归的变量筛选结果

变量	回归系数	标准误	Wald 卡方	自由度	P 值	OR 值
X_1(年龄)	0.057	0.022	6.790	1	0.009	1.059
X_2(胆红素)	0.402	0.105	14.799	1	0.000	1.495
X_3(白蛋白)	-1.715	0.536	10.234	1	0.001	0.180
常数项	1.757	2.364	0.553	1	0.457	5.796

1.BP 网训练集、校验集和测试集的确定

从原始数据中随机抽取 161 例作为训练集，53 例作为校验集，28 例为预测样本。

2. 输入数据的预处理

为使输入数据落在传递函数变化最快的梯度上，对输入数据进行预处理，用各指标的值除以该指标的最大值，即用 $x'_i=\frac{x_i}{\max(x)}$ 进行归一化处理。

3.BP 神经网络模型的建立及训练

网络输入层结点的个数由输入向量(自变量)的个数决定，网络选取 logistic 回归选出的 3 个变量作为输入。将病人生存时间按下式分为两类作为输出变量(应变量)$y_i(i=1,2,\cdots,n)$

1. 广东药学院预防医学系社会医学与卫生统计教研室(510224)
2. 山西医科大学

维普资讯 http://www.cqvip.com

$$y_i=\begin{cases}1 & \text{生存 } t\geqslant 5 \text{ 年}\\ 0 & \text{生存 } t<5 \text{ 年}\end{cases}$$

网络输出层为一个结点(患者存活大于5年,期望值定为1;存活小于5年,期望值定为0)。网络训练时采用了 Levenberg-Marquardt 优化算法,学习率为0.01,传递函数采用对数 S 形传递函数,隐单元数为2,使用 Matlab 5.3 软件。采用"早停止"策略防止过度拟合。

4.logistic 回归模型与神经网络模型在回代和前瞻性考核中的比较

我们把含有2个隐单元的 BP 网络和 Logistic 回归模型的拟合效果作比较,BP 网络方法回代的符合率为91%,logistic 回归为81%,(配对 $\chi^2=31.33$, $P<0.001$),BP 模型的判别效果不论在灵敏度、还是特异度上都优于 logistic 回归,见表2。

表2 不同界值时 BP 网络和 logistic 回归灵敏度、特异度比较

界值	BP 网络		logistic 回归	
	灵敏度(%)	特异度(%)	灵敏度(%)	特异度(%)
0.1	98.8	77	98.8	33
0.2	97	77	97	40
0.3	95	77	94	53
0.4	95	83	91	62
0.5	91	87	83	74
0.6	85	90	82	81
0.7	80	93	65	92
0.8	71	97	42	96
0.9	67	98	17	97

为了综合比较不同分类界值情况下 BP 网络和 logistic 回归的灵敏度和特异度,采用 ROC 分析方法。

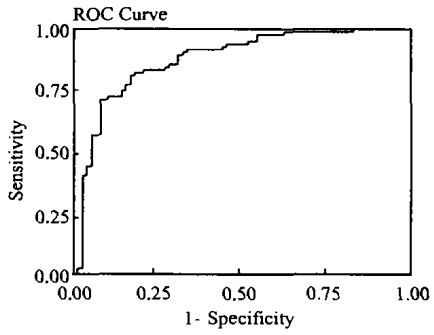

图1 logistic 回归模型判别结果的 ROC 曲线

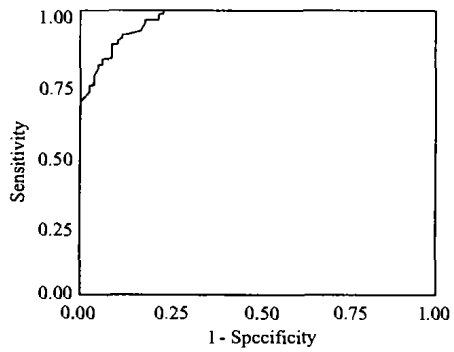

图2 BP 神经网络判别结果的 ROC 曲线

logistic 回归模型的 ROC 曲线下面积 $A_{Z1}=0.879$,标准误 $SE_1=0.028$;BP 网络的 ROC 曲线下面积 $A_{Z2}=0.97$,标准误 $SE_2=0.010$。两个 ROC 曲线下面积有显著性差异,($Z=3.652$, $P<0.05$)。

为了检验所建模型是否合适,我们对28个预测数据测试样本进行测试,BP 模型预测全部正确(100%),而 logistic 回归有3个错判,正确率为89%,由此可见,BP 模型不论在拟合还是预测方面都取得很好的效果。为了进一步说明模型的预测能力,应增加需要预测的新样本数。

讨 论

1.BP 神经网络目前在统计学领域主要应用在预测、判别分类问题中,它通过对有代表性实例的学习和训练,能够掌握事物本质特征,一个训练好的 BP 网络,理论上能够逼近任何输入(自变量)和输出(应变量)之间的任意非线性映射,具有很强的自组织、自适应能力,有高度的容错性。该模型在应用中自变量可以是连续的,也可以是离散的,不需要考虑自变量是否满足正态性及变量间独立等条件,可以识别变量间复杂的非线性关系,尤其是用现有统计方法无法达到目的或效果不好时,采用此模型往往收到很好的效果。而传统的判别分析往往对数据的分布有各种假设条件的要求,例如 Fisher 判别要求自变量必须是数值型的,Bayes 判别要求资料服从多变量正态分布,用以判别的自变量相互间要独立等等;不满足这些条件时,可能要对原始数据作变量变换(包括如何使非线性关系变换为线性关系),而实际上选择哪一种函数变换是很困难的,可能要用到较复杂的统计方法。而 BP 网作为一个非线性的数学模型,在这些方面是有优势的,它尤其善于处理复杂模糊的映射关系,不需要知道数据的分布形式。BP 神经网络作为一个标准的非线性数学模型,在统计学有广阔的应用前景。

2.实际应用中使用 BP 网时,网络的设计至关重要。一般应从网络的层数、每层中神经元个数、传递函数的选择、权值、学习率的设置等几方面来考虑。实例分析中,要想进一步提高该方法的准确性,稳妥的方法是增加样本例数。使用的数据越多越全面,则其中所隐含的事物本身规律就越强,利用 BP 网从中所抽取的函数关系就越具有普遍性,因而就更准确。网络累计足够的信息,就能建立一个智能系统来帮助临床医师,尤其是无经验的临床医师进行辅助诊断和预后分析[3~4]。

3.BP 网络应用于统计领域尚有一些问题有待解决,如权重系数的假设检验,计算权重系数的可信区间,含隐含层时权重系数的医学解释,判断输入变量的判别能力,输入变量的选择等都还需要进一步研究。

维普资讯 http://www.cqvip.com

The Research about the Comparison of Neural Network Model with logistic Regression *Li Lixia, Wang Tong, Fan Fengxi. Department of Medical Statistics, Guang Dong Pharmacy University*(510224), *Guangzhou*

【Abstract】 Objective To explore the application of BP neural network on discriminant analysis through comparing with logistic regression model. **Methods** Levenberg-Marquardt algorithm is adopted which makes learning time short, convergence fast, early-stopping method is used for avoiding over-fitting. And compare the performance of a neural network model with that of logistic regression model. **Results** BP neuralnetwork gets good results in internal validation and external validation, the comparison of their ROC curves(relative operating characteristic curve) also give a good prove. **Conclusion** BP neural network is worthy to be popularized, especially in the fields of survival analysis and discriminant analysis.

【Key words】 BP neural network; Over-fitting; BP algorithm; ROC curve; Logistic regression

参 考 文 献

1. 余雪丽主编. 神经网络与实例学习. 中国铁道出版社, 1996.
2. 薛禾生. 人工神经网络方法. 中国医院统计, 1999, 6(2): 100-102.
3. Mango LJ. Computer-assisted cervical cancer screening using neural networks. Cancer Letter, 1994, 77: 155-162.
4. Edwards F, Zazulia AR. Artificial neural networks improve the prediction of mortality in intracerebra hemorrhage. Neurology, 1999, 53: 351-357.

谈谈医院床位利用指标的两种不同计算口径

浙江省余姚市人民医院(315400)　李优军

医院的床位利用情况主要是通过"病床使用率","病床周转次数"等指标来反映的。"病床使用率"与"病床周转次数"均与实际开放总床日数有关,目前各医院在计算该两项指标过程中,对实际开放总床日数的计算口径存在着两种不同的理解。

一种是将实际开放总床日数理解为编制床位开放总床日数,把编制床位以外的开设床不管是固定的,还是短期内就要撤除的,均作为临时加床;另一种是将实际开放总床日数理解为固定床位开放总床日数,而只把紧急处置病人而增设的病床,并在短期内即要撤除的作为临时加床。

由于医院改革的深入,绝大多数医院在将社会效益摆在医院工作首位的同时,也注重不断提高经济效益,多数医院的固定病床数都大于编制床位数。所以单一的用上述病床使用指标的两种计算方法的一种来反映医院病床的使用情况是不够全面的,应用两种计算口径分别算得的床位使用指标来同时反映医院病床在一定时期的负荷情况,这样就能从两个不同的角度为医院管理提供病床使用方面的资料,既反映了编制病床的使用情况,又反映固定床的使用情况。这里值得一提的是,编制床位的确定具有一定的科学性。因此,医院在增设固定床位时,要根据各科的医护力量及医院的管理水平而定,不应盲目增设,也不应得过且过。当用第一种方法算得的指标反映出床位超负荷时,应注意医疗质量的管理,在发现医疗质量下降时,并确系床位超负荷引起的,就应适当减少床位负荷,提高医疗质量。反之,当用第二种方法所算得的指标反映出床位满足不了病人需要时,应扩大医院面积,增加医院固定床位,保障人民的身心健康。

由于我院目前仍只单一使用上述第二种计算方法来计算床位使用指标,缺乏一定的对比资料。现将某医院的资料进行对比如下:某医院 1987 年编制床位为 400 张,平均开放固定床位 456 张。出院人数为 10 920 人,实际占用总床日数为 160 095 日,该医院 1987 年病床利用情况如表 1。

表 1　某院 1987 年病床利用情况

	病床周转次数	病床工作日	病床使用率(%)
编制床位	27.3	400.20	109.70
固定床位	23.9	351.10	96.20

按二级医院标准,平均病床周转次数不低于 22 次,按此标准就固定床位利用情况来看,该院平均病床周转次数刚刚出现超负荷现象,但按编制床位利用情况看,该院病床周转次数大大超过 22 次,出现超负荷现象。因为对二级医院要求是病情难度大、病人住院天数长、病床周转次数略低于一级医院。所以该院工作量出现超负荷现象,应特别注意医疗质量的管理。

为适应我院深化改革的需要,建议我院应用两种计算口径,分别计算床位使用指标,这样就能从不同角度反映出医院病床在一定时期内的负荷情况。例我院 2002 年 1～12 月份编制床位为 440 张,平均开放固定床位为 516 张,出院人数为 15 610 人,实际占用总床日数为 180 126 日,我院 2002 年 1～12 月份病床利用情况如表 2。

表 2　我院 2002 年病床利用情况

	床位周转次数	床位工作日	病床使用率(%)
编制床位	35.48	409.38	112.16
固定床位	30.25	349.08	95.64

结合浙江省最新医院评审标准三级医院实际病床使用率:90%≤三级医院≤105%。可见我院按编制床位算得的病床使用率反映出我院床位大大出现超负荷现象。按固定床位算得的病床使用率反映出我院床位尚未达到饱和状态。所以建议我院应提高医院的医技质量及医院的管理水平,改善医院的服务态度,充分发挥医院现有的仪器设备作用,调动员工的积极性和创造性,使我院永远立于不败之地。