·基础理论与方法·

人工神经网络在失去土地农民心理健康调查中的应用

张训保 黄水平 卓朗 吴秀娟 孙桂香 赵华硕 李磊

【导读】 介绍人工神经网络(ANN),结合实例比较与 logistic 回归在解决分类问题的优缺点。以1070名失去土地农民心理健康调查资料为例建立 ANN 模型与 logistic 回归模型,比较两种模型的优劣。测试集样本 BP 神经网络预测精度为94.299%,logistic 回归预测精度为51.028%,BP 神经网络具有良好的泛化能力。结论:当传统统计分析条件不能得到满足或效果不佳时 ANN 能够达到良好的预测结果,在医学领域具有较好的应用前景。

【关键词】 人工神经网络; logistic 回归; 失去土地农民

The application of artificial neural network in studying landless farmer's mental health problems

ZHANG Xun-bao, HUANG Shui-ping, ZHUO Lang, WU Xiu-juan, SUN Gui-xiang, ZHAO Hua-shuo, LI Lei. Department of Medical Statistics and Epidemiology, School of Public Health, Xuzhou Medical College, Xuzhou 221002, China

【Introduction】 To introduce a method of classification with high precision — the artificial neural network (ANN), and to compare the results using logistic regression method. Using data from 1070 landless peasants' mental health survey, the artificial neural network models and logistic regression model were built and compared on their advantages and disadvantages of the two models. The prediction accuracy for artificial neural network was 94.229% and for logistic regression it was 51.028%. ANN appeared to have had good ability on generalization. ANN displayed advantages when conditions of classical statistical techniques could not be met or the predictive effect appeared to be unsatisfactory. Hence, ANN would make a better facture of its application in medical researche.

【Key words】 Artificial neural network; Logistic regression; Landless peasants

分类在医学研究中是一项非常重要的任务,旨在找出把给定数据集分成若干类的具体分类规则。人工神经网络(ANN)是目前最常用的解决分类规则提取问题的机器学习方法之一[1]。ANN 是数据挖掘中的一种高效的分类方法[2],它是由大量的处理单元互相连接而成的网络,能够模拟大脑的基本特性,进行识别和训练,而且不需要先验知识,但其内部规则可理解性差,不易从中提取规则。常用的 ANN 模型有前向网络的 BP 模型和径向基网络的 RBF 模型。根据两种算法的特点,引以具体的实例,将进一步对决策树和 ANN 算法性能进行对比分析,具体试验选用的是 ANN 中的 BP 网络模型和最常用的 logistic 回归模型,BP 网络模型是 ANN 的典型模型,而 logistic 回归是预测与因素筛选最常用的基本模型,用它们两者的对比分析可以很好地证实 ANN 与 logistic 回归分类预测性能的差别。

基本原理

ANN 是借鉴生物神经系统提出来的,其基本原理是神经元之间的信号传播与反馈[3]。其中常用的 BP 网络为三层结构,包含输入层、输出层以及处于输入输出层之间的隐层(图 1)。每层至少有一个以上结点(神经元),与层对应地输入结点、隐结点和输出结点[4]。隐层中的结点不直接与外界连接,但其状态影响输入输出之间的关系。输入结点、隐结点、输出结点之间的关系通过网络系数(权重和阈值)来反映。训练前,先通过一定方法赋予网络初始权重和阈值,然后利用训练样本对网络进行训练,使网络输出满足期望需求。训练成熟后的神经网络即可投入实际应用,可用于分类、预测等决策支持。BP 神经网络学习计算分为几个步骤:

1. 初始化网络系数,随机赋给各层结点之间的

基金项目:江苏省科技厅社会科学基金课题(BS2005018)

作者单位:221002 徐州医学院公共卫生学系

连接权重及阈值。

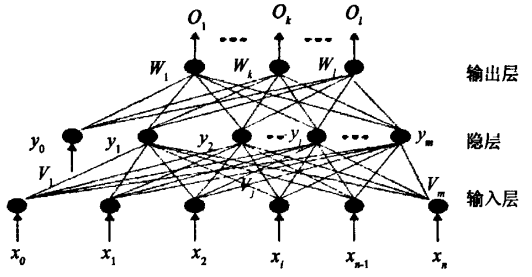

图1 BP 神经网络示意图

2. 输入训练样本向量($x_0, x_1, \cdots, x_j, \cdots$)。

3. 计算隐层和输出层结点的输出,其中①隐结点 i 的输出 y_i 用公式(1)计算:

$$y_i = f(\sum_j W_{ij}x_j - \theta_i) = f(net_i) \tag{1}$$

式中,y_i 为隐层第 i 个结点的计算输出,W_{ij} 为输入结点 j 至隐结点 i 的连接权重,x_j 为输入结点 j 的输入值,θ_i 为隐结点 i 的阈值,f 为激活(作用)函数,最常用激活函数为 S 型的 Sigmoid 函数。②输出结点 l 的输出 o_l 用公式(2)计算:

$$o_l = f(\sum_i T_{li}y_i - \theta_l) = f(net_l) \tag{2}$$

式中,o_l 为输出结点 l 的计算输出,T_{li} 为隐层结点 i 至输出结点 l 的连接权重,θ_l 为输出结点 l 的阈值,其他参数含义同前。

4. 计算误差:

(1)总误差 E 按公式(3)计算:

$$E = \frac{1}{2}\sum_l (t_l - o_l)^2 = \frac{1}{2}\sum_l [t_l - f(\sum_i T_{li}y_i - \theta_l)]^2$$
$$= \frac{1}{2}\sum_l \{t_i - f[\sum_i T_{li} f(\sum_j W_{ij}x_j - \theta_i) - \theta_l]\}^2 \tag{3}$$

式中,t_l 为输出结点 l 的期望输出值,其他参数含义同前。

(2)输出结点 l 的误差为 δ_l 根据公式(4)计算:

$$\delta_l = (t_l - o_l) f'(net_l) \tag{4}$$

(3)隐结点 i 的误差为 δ'_i 根据公式(5)计算:

$$\delta'_i = f'(net_i) \sum \delta_l T_{li} \tag{5}$$

5. 修正连接权重及阈值:

(1)隐结点 i 到输出结点 l 的连接权重按公式(6)修正:

$$T_{li(k+1)} = T_{li(k)} + \Delta T_{li} = T_{li(k)} + \eta\delta_l y_i \tag{6}$$

式中,$T_{li(k)}$ 为修正前的连接权重,η 为学习率,其他参数含义同前。

(2)输入结点 j 到隐结点 i 的连接权重按公式(7)修正:

$$W_{ij(k+1)} = W_{ij(k)} + \Delta W_{ij} = W_{ij(k)} + \eta'\delta'_i x_j \tag{7}$$

式中,$W_{ij(k)}$ 为修正前的连接权重,η' 为学习率,其他参数含义同前。

(3)输出结点 l 的阈值按公式(8)修正:

$$\theta_{l(k+1)} = \theta_{l(k)} + \eta\delta_l \tag{8}$$

式中,$\theta_{l(k)}$ 为修正前的阈值,其他参数含义同前。

(4)隐结点 i 的阈值按公式(9)修正:

$$\theta_{i(k+1)} = \theta_{i(k)} + \eta'\delta'_i \tag{9}$$

式中,$\theta_{i(k)}$ 为修正前的阈值,其他参数含义同前。

6. 转步骤 2,直到收敛。

实例分析

随着我国工业化、城市化进程不断加快,城市现有的土地已经不能满足高速发展的经济需要,于是城市逐渐向周边农村扩张[5],大批农民离开祖祖辈辈赖以生存的土地,被冠以"失地农民"的标志,徘徊在城市和农村的交界处,心理出现不适应及发生心理异常[6]。本调查于 2007 年 12 月对徐州市周边地区 1070 名失地农民开展调查,内容包括基本情况、生活习惯、社会关系、健康状况、社会关怀等 63 项及心理状况评定量表(SCL-90),了解失地农民心理健康状况,观察影响其心理健康的因素,为制定失地农民的相关政策提供决策依据。

1. 变量选择:Epi Data 3 软件建立数据库后同时建立 BP 神经网络、logistic 回归模型。采用 SPSS 公司 Clementine 12.0软件建立模型,以比较两种模型的优劣。数据审核之后,建立缺失值超节点以 C&RT 算法填补缺失值。两种模型均选择全部 63 项指标,将 1070 名失地农民随机分为 70% 训练样本集、30% 测试样本集。

2. 建立模型:

(1)BP 神经网络模型建立:先后采用 1、2、3 隐层数建立模型,比较三种结构的拟合效果(图 2)。从图 2 可以看出 BP 网络不随隐层数增加而精度增加,故

选择单隐层 BP 网络。初始学习率设置为0.3，为提高训练速度在模型训练中增加动量项α=0.9。

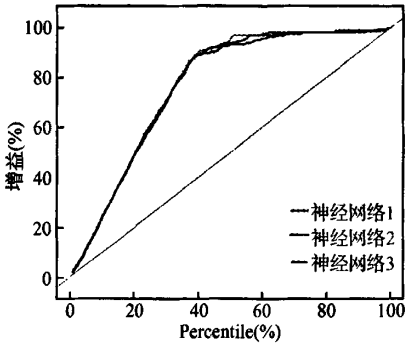

图2 三种不同结构 BP 网络收益图

(2) logistic 回归模型：采用二分类非条件 logistic 回归建立模型，选入方程内变量的方法为全变量模型。建立好的模型其预测精度为51.028%。

3. 结果分析：从图 3 可以看出，BP 神经网络在前期快速升高并在达到高点后趋于平稳，logistic 回归增益上升较慢。图 3 从左至右的走势通常是从 0% 到 100%。优秀模型的收益图将陡升至 100%，然后保持平直。无法提供有用信息的模型将呈对角线状，即从左下角到右上角。由此看来 BP 神经网络拟合效果较优。

经测试集测试结果如表 1 所示，BP 神经网络预测精度为 94.299%，ROC 曲线下面积为 0.962；logistic 回归预测精度为51.028%，ROC 曲线下面积为0.771，两模型建模时间均<1 min。可以看出 BP 神经网络具有较好的泛化能力。

为观察各模型中自变量对应变量的影响程度，输出两种模型按因素重要性排序前 10 位因素顺位图(图 4)。由此可以看出，BP 神经网络与 logistic 回归筛选出的前 10 位因素可概括为 3 个层次：个体层次、社会支持层次、行为层次。而 logistic 回归只能说明各变量的流行病学意义，是 BP 神经网络模型不能及的特性。

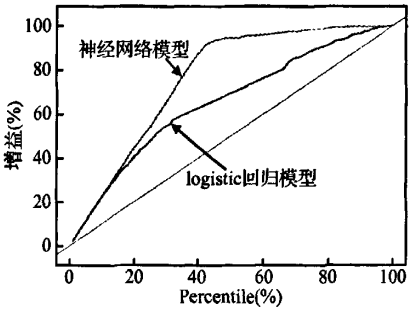

图3 两种模型收益图

表1 两种模型测试结果比较

模型	建模时间(min)	预测精度(%)	ROC 曲线下面积
BP 神经网络	<1	94.299	0.962
logistic 回归	<1	51.028	0.771

讨 论

从预测角度来看，ANN 预测精度较 logistic 回归精度高。这种高精度易实施的模型在将来的心理健康筛检中可以考虑采用。综合两模型因素顺位可见危险因素分为 3 个层次：个体层次、社会支持层次、行为层次。从个体层次上，中青年、健康状况不好的容易产生敌对心理，而与是否拥有医疗保险无关，提示敌对心理可能与个体心理期待和欲求较高而现实的满足程度较低有关，简单的医疗服务可能无能为力。从社会支持层次上父母身体状况好、对政府的补助满意、能够得到家人的帮助不容易产生敌对心理。行为层次上，经常与别人聊天，对未来充满信心成为保护因素，一般的娱乐休闲活动没有影响敌对因子。

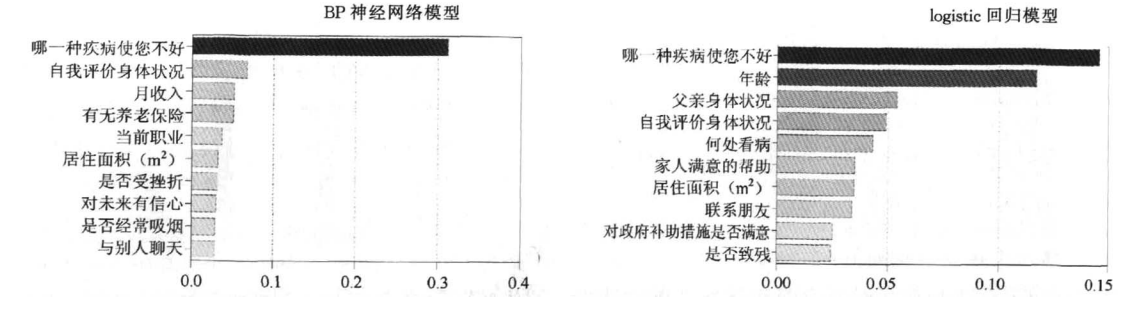

图4 BP 神经网络和 logistic 回归影响因素顺位图

目前 ANN 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，尤其是在用现有统计方法无法达到目的或效果不好时，与现有方法相比，优点显而易见，其应用潜力有待于广大学者共同开发。

从国内外其他一些医学应用实例看，ANN 处理医学问题的能力显而易见，不同的方法在不同的资料中优点各异。ANN 预测精度方面优于 logistic 回归，而在变量筛选方面不及 logistic 回归解释性强，在实际医学资料中，如果能将 ANN 与其他方法联合应用，充分利用各种方法的优点，将大大提高各模型的应用价值。

此外，ANN 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以及网络结构的优化选择等都还需要进一步研究。

参 考 文 献

[1] Butler R, Welch V, Engert D, et al. A national-scale Authentication Infrastructure. IEEE Computer, 2000, 33(12): 60-66.

[2] Mehmed Kantardzic. Data Miuing Concpts, Models, Methods, and Algorithms(数据挖掘概念、模型、方法和算法). 闪四清，陈茵，程雁，等译. 北京：清华大学出版社，2003：171-195.

[3] 韩力群. 人工神经网络教程. 北京：北京邮电大学出版社，2006：29-36.

[4] 钱玲，施侣元，程茂金. 人工神经网络应用于糖尿病/糖耐量异常的疾病分类研究. 中华流行病学杂志，2003，24(11)：1052-1056.

[5] 涂文明. 城市化进程中失地农民社会保障模式的选择和构建. 理论导刊，2004，12：33-35.

[6] 赵翠玲，李明蔚. 失地农民社会心理与社会稳定刍议. 山东省农业管理干部学院学报，2007，23(1)：18-19.

(收稿日期：2008-05-06)

(本文编辑：张林东)

·读者·作者·编者·

本刊对统计学方法的要求

研究设计：应告知研究设计的名称和主要方法。如调查设计(分为前瞻性、回顾性还是横断面调查研究)，实验设计(应告知具体的设计类型，如自身配对设计、成组设计、交叉设计、析因设计、正交设计等)，临床试验设计(应告知属于第几期临床试验，采用了何种盲法措施等)；主要做法应围绕 4 个基本原则(重复、随机、对照、均衡)概要说明，尤其要告知如何控制重要非试验因素的干扰和影响。

资料的表达与描述：用 $\bar{x} \pm s$ 表达近似服从正态分布的定量资料，用 $M(Q_R)$ 表达呈偏态分布的定量资料，用统计表时，要合理安排纵横标目，并将数据的含义表达清楚；用统计图时，所用统计图的类型应与资料性质相匹配，并使数轴上刻度值的标法符合数学原则；用相对数时，分母不宜小于 20，要注意区分百分率与百分比。

统计学分析方法的选择：对于定量资料，应根据所采用的设计类型、资料具备的条件和分析目的，选用合适的统计学分析方法，不应盲目套用 t 检验和单因素方差分析；对于定性资料，应根据所采用的设计类型、定性变量的性质和频数所具备的条件及分析目的，选用合适的统计学分析方法，不应盲目套用 χ^2 检验。对于回归分析，应结合专业知识和散布图，选用合适的回归类型，不应盲目套用直线回归分析；对具有重复实验数据检验回归分析资料，不应简单化处理；对于多因素、多指标资料，要在一元分析的基础上，尽可能运用多元统计分析方法，以便对因素之间的交互作用和多指标之间的内在联系做出全面、合理的解释和评价。

统计结果的解释和表达：当 $P<0.05$(或 $P<0.01$)时，应说对比组之间的差异具有统计学意义，而不应说对比组之间具有显著性(或非常显著性)差异；应写明所用统计分析方法的具体名称(如：成组设计资料的 t 检验、两因素析因设计资料的方差分析、多个均数之间两两比较的 q 检验等)，统计量的具体值(如：$t=3.45$，$\chi^2=4.68$，$F=6.79$ 等)；在用不等式表示 P 值的情况下，一般情况下选用 $P>0.05$、$P<0.05$ 和 $P<0.01$ 三种表达方式即可满足需要，无须再细分为 $P<0.001$ 或 $P<0.0001$。当涉及总体参数(如总体均数、总体率等)时，在给出显著性检验结果的同时，再给出 95% 可信区间。

本刊编辑部