

小样本量大肠癌数据对 BP 神经网络准确度的影响探讨^{*}

刘秀峰 刘 芬 李金城

(广州中医药大学医学信息工程学院 广州 510006)

〔摘要〕 采用二进制编码对数据进行量化,建立基于 BP 神经网络的大肠癌证型分类模型并分析其优势,探索 8 个证型的样本量对 BP 神经网络准确度的影响,通过删减较少样本的大肠癌证型得出样本量与准确度的量化关系。

〔关键词〕 大肠癌;小样本量;BP 神经网络;准确度

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2018.08.013

Discussion on the Impact of Colorectal Cancer Data of a Small Sample Size on BP Neural Network Accuracy LIU Xiu-feng, LIU Fen, LI Jin-cheng, College of Medical Information Engineering, Guangzhou University of Chinese Medicine, Guangzhou 510006, China

〔Abstract〕 The paper quantifies data by using binary coding, builds the syndrome type classification model of colorectal cancer based on BP neural network, analyzes its advantages, and explores the impact of the sample size of 8 syndrome types on Back Propagation (BP) neural network accuracy. By ruling out colorectal cancer syndrome types of a small number of samples, it gains the quantified relation between sample size and accuracy.

〔Keywords〕 Colorectal cancer; Small sample; BP neural network; Accuracy

1 引言

大肠癌是临床上常见的消化道恶性肿瘤之一,包括结肠癌、直肠癌和肛管癌^[1]。大肠癌治疗的主要手段是手术切除以及辅助放化疗,但其毒副作用影响患者治疗效果及生活质量。研究表明在大肠癌治疗方面中医药具有一定的优势^[2],大肠癌证型分

类标准能一定程度上促进中医药大肠癌治疗。经研究发现基于误差逆传播(Back Propagation, BP)神经网络的分类模型较决策树分类模型更擅长于非线性映射关系的处理^[3]。基于临床数据较难获取等客观条件,本研究在上述基础上研究小样本量的大肠癌数据对 BP 神经网络准确度的影响,探索用 BP 神经网络方法训练大肠癌证型分类模型时至少需要多少样本量。这对其他癌症研究应用 BP 神经网络方法具有一定的借鉴意义,且更能贴近现实条件。

〔修回日期〕 2018-03-23

〔作者简介〕 刘秀峰,教授,发表论文 20 余篇。

〔基金项目〕 广州中医药大学薪火计划(项目编号: XH20160105)。

2 数据收集及处理

2.1 数据来源

本研究共收集 338 例大肠癌患者数据，其中 188 例来源于中国知网全文数据库、维普全文数据库及万方数据知识服务平台的文献数据，150 例来源于广东省中医院临床案例。筛选出符合标准的 218 例，其中男 108 例，女 110 例，年龄为 25 ~ 80 岁，平均年龄为 61 岁。

2.2 纳入、排除标准

纳入标准：经细胞学或病理学检查诊断为结直肠癌；大肠癌术后（经手术切除原发病灶、术后病程 3 个月内）；年龄在 25 ~ 80 岁；检查资料完整。排除标准：病理诊断不明确；合并有严重的心、肺、脑、肝、肾等疾病；因其他原因 3 个月内死亡；多原发性肿瘤或既往有恶性肿瘤病史。

2.3 数据量化

筛选出大肠癌患者所表现出的 28 项体征，包括里急后重、大便秘结、脉沉、舌红、舌苔厚等症状。症状量化标准依据《中医症状鉴别诊断学》（第 2 版）^[4]，通过该书对症状的分类进行初步划分。规范症状时将已确定的大肠癌常见症状进行赋值量化，有该症状者为 1，无该症状者为 0。依据《中华肿瘤治疗大成》^[5]、《中华中医药学会标准肿瘤中医诊疗指南》^[6] 及临床经验进行大肠癌证型术语规范化，将辨证结果进行比较分析，再与专家进行讨论，最终形成一致意见。初步将多种证型术语规范为 8 种并进行赋值量化，有该证型者为 1，无该证型者为 0。样本证型分布，见表 1。

表 1 大肠癌证型分布

证型	例数	百分比（%）
脾失健运	65	29.82
脾虚夹瘀	39	17.89
湿热内蕴	36	16.51
气滞血瘀	20	9.17
瘀毒内阻	19	8.72
气血亏虚	17	7.80
脾肾阳虚	16	7.34
肝脾不调	6	2.75
合计	218	100.00

3 方法及结果

3.1 BP 神经网络

BP 算法是迄今最成功的神经网络学习算法^[7]，将 BP 算法应用于多层前馈神经网络可称为误差逆传播神经网络，是通过样本数据的训练，不断修正网络权值和阈值使误差函数沿负梯度方向下降，逼近期望输出的一种应用较为广泛的神经网络模型，多用于函数逼近、模型识别分类、数据压缩和时间序列预测等探究与建模，具有高度非线性和较强的泛化能力。

3.2 BP 神经网络在大肠癌证型分类中的优势

3.2.1 全面提高特异性与准确度 利用神经网络并行结构和并行处理的特征，通过适当选择评价项目能全面提高特异性与准确度。吴拥军等通过神经网络研究肺癌诊断^[8]总结出神经网络处理优于常规统计学方法，在特异性和准确度上有明显差别，其结果分别较常规统计学方法提高 28.0% 和 21.4%。

3.2.2 实现系统模糊评价 薛佳殷^[9]等通过建立 BP 神经网络，选取吸烟、饮用自来水、喜食豆制品、常生气、年龄和生活节奏快等指标对是否为胃癌高危人群进行分类，取得良好效果。该网络强大的容错性可以实现对大肠癌系统的模糊评价。

3.2.3 分类效果更佳 孙文恒^[10]等研究胰腺癌诊断发现该网络具有一定的泛化能力，对于未学习的样本同样具有很好的分类能力，结果表明该网络对胰腺癌预测效果较好。所以利用该特点将在一定程度上使大肠癌证型分类效果更为合理。

3.3 结构及隐层设置

本研究选用的 BP 神经网络由输入层、隐藏层及输出层 3 层单元组成。输入层由主成分分析得出的 22 个主成分所决定；输出层由 8 个证型指标决定。根据有关研究可知如果有 1 个隐层的神经网络，只要隐节点足够多就可以利用任意精度逼近 1 个非线性函数。故建立基于含有 1 个隐层的 3 层多输入、输出的 BP 网络预测模型。在整个网络设计

过程中隐层神经元个数的确定极为重要, 过多会加大网络计算量且容易产生过度拟合; 过少则会减弱网络性能, 达不到预期效果。经验表明网络中隐层神经元的数目与实际问题的复杂程度、输入和输出层的神经元数以及对期望误差的人为设定都有直接的联系, 所以目前对于隐层中神经元数目的确定并没有明确的公式, 需要根据经验和多次实验来最终确定。本研究在选取隐层神经元个数时参照经验公式 $\sqrt{n+m+a}$, 其中 n 为输入层神经元个数, m 为输出层神经元个数, a 为 $[1, 10]$ 之间的常数。BP 神经网络结构, 见图 1。

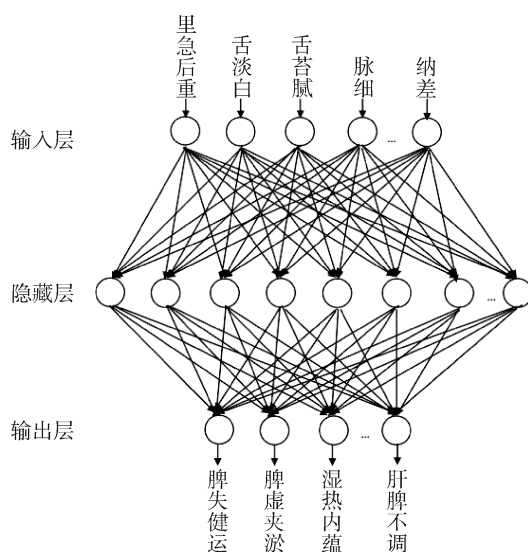


图 1 BP 神经网络结构

3.4 输出方式

将 8 个证型的样本期望输出值设为 $(1, 0, 0, 0, 0, 0, 0, 0)$, $(0, 1, 0, 0, 0, 0, 0, 0)$, $\dots$, $(0, 0, 0, 0, 0, 0, 0, 1)$, 隐层与输出层之间的激活函数采用正切 S 型传递函数 $\tanh$ 。参考 BP 神经网络与决策树的大肠癌虚实证型分类对比研究中的阈值^[3], 在观察测试集数据的输出值后依据正

确率进而不断调试区间, 确定将预测结果 ≥ 0.65 设为诊断成立。

3.5 3 倍交叉验证过程

3.5.1 第 1 个 BP 神经网络模型 将主成分分析后的数据导入 Python 数据框中, 将数据集分成 3 部分, 进行 3 倍交叉验证, 通过不断调试步长与迭代次数, 使网络的系统误差达到最小且该神经网络趋于收敛。在步长为 0.05、迭代次数为 50 000 时网络性能达标, 训练自动停止。建立第 1 个 BP 神经网络模型, 其测试结果, 见表 2。可看出样本分布不均匀; 数量少的样本容易对网络训练造成干扰。以脾虚夹瘀证为例, 第 1 次交叉验证的测试例数为 0, 这是由样本分布不均匀造成。第 2 次交叉验证的测试正确率偏高的原因可能与前面训练集的数据特征比较集中有关。而第 3 次交叉验证的测试正确率偏低的原因可能是训练集特征不集中或是其他证型对该证型的判别过程产生干扰。

3.5.2 第 2 个 BP 神经网络模型 为解决这些问题提出的方法是: 添加随机函数, 打乱样本; 删除少于 5% 的样本量的证型 (即肝脾不调证)。建立第 2 个 BP 神经网络模型, 发现即使删除 1 个证型, 该模型的步长和迭代次数无变化。达到模型收敛后各模型的测试结果, 见表 3。该举措小幅度地提高网络的测试判别正确率, 一定程度上证实数量少的样本容易对网络训练造成干扰。但由于训练时有些样本较少所以测试时正确率偏低, 且各证型对应的特征不太明显。以气滞血瘀证为例, 3 次交叉验证的测试结果均为 25%, 虽然较第 1 个模型有所提升, 可能由于其他证型的干扰导致该测试结果不太理想。通过观察气滞血瘀证与瘀毒内阻证的样本量相近, 结果差异较大的原因可能与样本特征值集中与否有关。

表2第1个BP神经网络模型交叉验证结果

证型	第1次交叉验证			第2次交叉验证			第3次交叉验证		
	测试例数	正判例数	正确率(%)	测试例数	正判例数	正确率(%)	测试例数	正判例数	正确率(%)
脾失健运	22	12	54.55	19	10	52.63	24	15	62.50
脾虚夹瘀	0	0	0	20	15	75.00	19	8	42.11
湿热内蕴	8	4	50.00	11	7	63.64	17	6	35.29
气滞血瘀	5	2	40.00	12	0	0	3	0	0
瘀毒内阻	15	0	0	3	0	0	1	0	0
气血亏虚	7	0	0	4	0	0	6	0	0
脾肾阳虚	10	0	0	4	0	0	2	0	0
肝脾不调	5	0	0	0	0	0	1	0	0
平均正确率(%)	-	-	25	-	-	43.84	-	-	39.73

表3第2个BP神经网络模型交叉验证结果

证型	第1次交叉验证			第2次交叉验证			第3次交叉验证		
	测试例数	正判例数	正确率(%)	测试例数	正判例数	正确率(%)	测试例数	正判例数	正确率(%)
脾失健运	24	11	45.83	22	16	72.73	19	11	57.89
脾虚夹瘀	13	9	69.23	15	8	53.33	11	7	63.64
湿热内蕴	11	6	54.55	10	4	40.00	15	7	46.67
气滞血瘀	8	2	25.00	8	2	25.00	4	2	50.00
瘀毒内阻	7	0	0	5	0	0	7	0	0
气血亏虚	4	0	0	3	0	0	10	0	0
脾肾阳虚	4	1	25.00	7	0	0	5	0	0
平均正确率(%)	-	-	40.85	-	-	42.86	-	-	35.21

3.5.3 设计不同的神经网络模型验证观察结论类似于第2个模型的做法，首先删除脾肾阳虚证，使用6个证型做模型；继而删除气血亏虚证，使用5个证型做模型；最后删除瘀毒内阻证，使用4个证型做模型。通过分析各证型的测试正确率，发现删除完瘀毒内阻证后使用4个证型做模型时，该测试结果的正确率得到大幅度的提高。该过程同样添加随机函数，打乱样本。最后1个模型的步长为0.035，迭代次数为20 000时网络达到收敛。最后1

个BP神经网络模型的测试结果，见表4。可知单证的测试正确率最高可达86.96%，3次交叉验证的平均测试正确率达到74.37%。正确率大幅度的提高从一定程度上表明其他样本量少且特征值不集中的证型会对其判断造成一定的干扰。在构建BP神经网络模型的过程中，训练样本集的样本量尽量达到10%以上可以保证一定的测试正确率。当所占百分比达到30%左右，该证型的测试正确率可达85%以上。

表 4 最后 1 个 BP 神经网络模型交叉验证结果

证型	第 1 次交叉验证			第 2 次交叉验证			第 3 次交叉验证		
	测试例数	正判例数	正确率 (%)	测试例数	正判例数	正确率 (%)	测试例数	正判例数	正确率 (%)
脾失健运	22	19	86.36	23	20	86.96	20	17	85.00
脾虚夹瘀	13	10	76.92	15	12	80.00	11	9	81.82
湿热内蕴	10	8	80.00	9	4	44.44	17	11	64.71
气滞血瘀	8	3	37.50	6	3	50.00	6	3	50.00
平均正确率 (%)	-	-	75.47	-	-	73.58	-	-	74.07

4 结论

基于黑箱结构的人工神经网络能利用其自主学习能力^[11-12]，用 BP 神经网络建模可将过程或对象看作是 1 个“黑箱”，只要确定输入和输出就可以建立相应的模型，不必像传统的系统辨识那样将过程辨别为线性还是非线性，这有利于对未知过程的系统进行建模。本研究通过构建 BP 神经网络，探索小样本量大肠癌数据对 BP 神经网络准确度的影响得出以下结论：（1）删除少于 5% 样本量（本研究为 6 例）的证型，该模型的步长和迭代次数不变。（2）样本量少于 5% 时一般不建议进行 BP 神经网络构建，不仅可能对其他判断有影响，该样本也不能做出很好的拟合。（3）当样本量占总样本的 10% 左右（本研究为 20 例）且该特征值比较集中时，能达到 50% 左右的测试正确率。（4）当所占百分比达到 30% 左右（本研究为 65 例），该证型的测试正确率可达 85% 以上。（5）在进行 BP 神经网络训练时，样本变量不需要很多，因为神经网络的信息存储能力有限，过多的样本会造成一些有用的信息被丢弃。如果样本数量过多，应增加隐层节点数或隐层数目才能增强学习能力。

由前期的基础研究可知 BP 神经网络的分类模型较决策树分类模型更擅长于非线性映射关系的处理^[3]。本研究旨在探索小样本量对 BP 神经网络准确度的影响，通过定量分析得出相应的准确度。样本量对神经网络的影响不容忽视，特别是小样本量的临床癌症数据。虽然处于大数据时代，但未来一段时间内医学临床数据的整合仍是一个严峻的课题。

参考文献

1 司富春, 岳静宇. 近 30 年大肠癌中医证型和用药规律分析 [J]. 中华中医药杂志, 2012, 27 (7): 1929 - 1931.

2 陈叶, 刘金涛, 朱源, 等. 大肠癌中医辨证及治疗概况 [J]. 中国肿瘤. 2015, 24 (4): 319 - 324.

3 李金城, 从瑶, 陈秋芬, 等. 基于 BP 神经网络与决策树的大肠癌虚实证型分类对比 [J]. 医学信息学杂志, 2017, 38 (5): 61 - 64, 84.

4 姚乃礼. 中医症状鉴别诊断学 [M]. 北京: 人民卫生出版社, 2000: 27 - 35.

5 潘敏求. 中华肿瘤治疗大成 [M]. 石家庄: 河北科学技术出版社, 1996: 489 - 490.

6 中华中医药学会. 中华中医药学会标准: 肿瘤中医诊疗指南 [M]. 北京: 中国中医药出版社, 2008: 26 - 29.

7 周志华. 机器学习 [M]. 北京: 清华大学出版社, 2016.

8 吴拥军, 吴逸明, 张振中, 等. 基于人工神经网络的“最优标志物群”在肺癌诊断中的应用研究 [J]. 实用肿瘤杂志, 2002, 17 (5): 317 - 320.

9 薛佳殷, 何帅, 李苹, 等. 应用人工神经网络筛查胃癌高危人群的研究 [J]. 现代预防医学, 2016, 43 (2): 197 - 199.

10 孙文恒, 王炜, 周文策. 人工神经网络技术在胰腺癌诊断中的应用 [J]. 兰州大学学报 (自然科学版), 2008, 44 (S1): 224 - 227.

11 Dayhof JE, Deleo JM. Artificial Neural Networks [J]. Cancer, 2001, 91 (8): 1615 - 1634 .

12 Cross SS, Harrison RF, Kennedy RL. Introduction to Neural Networks [J]. Lancet, 1995, 346 (8982): 1075 - 1079.