三种统计学模型在糖尿病个体患病风险预测中的应用*

蚌埠医学院预防医学系(233030)　宋　健　吴学森△　张　杰　张玉媛　陈　雪

【提　要】　目的　探讨 logistic 回归、BP 神经网络和决策树分析模型在预测个体 2 型糖尿病患病风险中的应用。**方法**　分别应用 logistic 回归、BP 神经网络与决策树建立 2 型糖尿病预测模型,通过受试者工作特征曲线评价模型的预测效能。**结果**　共 550 名糖尿病患者和 1100 名非糖尿病患者纳入本次研究。logistic 回归、BP 神经网络和决策树分析模型的预测一致率分别为 80.8%、84.1% 和 81.1%。3 种模型 ROC 曲线下面积(AUC)分别为 0.739、0.777 和 0.737。BP 神经网络的 *AUC* 与 logistic 模型和决策树分析模型的均有统计学差异($P<0.05$)。**结论**　BP 神经网络在预测个体患 2 型糖尿病方面具有更好的预测效能。

【关键词】　2 型糖尿病　logistic 回归　BP 神经网络　决策树分析

2 型糖尿病是严重危害人类健康的重大公共卫生问题,全世界约有超过 3.5 亿人患有 2 型糖尿病[1]。中国是世界上糖尿病患病人数最多的国家,患病率高达 11.6%[2]。有效地对个体进行糖尿病风险评估,可以筛选出高危人群,并通过一系列的行为和生活方式干预,减少糖尿病及相关并发症的发生。数据挖掘技术是近些年来广泛应用于医学领域的一种新的分析方法,在疾病诊断、预后、风险评估等方面具有良好的应用价值[3-5]。数据挖掘技术可以充分利用已有数据的信息,从具有重复性、多样性及不规范性等特点的复杂的医学数据中提取出有价值的信息,并为临床决策提供帮助[6-7]。其中,应用最广泛的有采用误差反向传递(back propagation,BP)学习方法的 BP 神经网络和决策树分析模型。本文采用慢性病社区调查数据,探讨 BP 神经网络与决策树分析模型在糖尿病个体风险预测中的应用价值,并与传统的 logistics 回归进行比较,以求寻找到 2 型糖尿病风险预测的最佳数学模型。

资料与方法

1. 资料来源

本课题组于 2015 年 7 月至 8 月,采用横断面调查方法,选择蚌埠市龙子湖区共 7 个社区,以家庭为抽样单位,共收回有效问卷 3077 份。调查内容包括两个方面:问卷调查及体格和实验室检查。采用自行设计问卷,由经过培训的课题组成员对社区居民进行问卷调查。调查问卷信息主要包括:受访者的一般个人及家庭信息及生活行为方式;体格检查指标有身高、体重、腰围等;实验室检查指标主要包括:空腹血糖、血脂和糖化血红蛋白等。数据首先录入到 Epidata 软件中,采用双录入方式,并逐一核对。

2. 相关变量及定义

(1)体质指数(body mass index,BMI)= 体重(kg)/身高(m)^2,正常值:$18.5 \leqslant BMI < 24$,< 18.5 或者 > 24 均视为不正常;(2)高血压:收缩压/舒张压 ≥ 140/90mmHg 和/或已确诊为高血压者;(3)甘油三酯:正常值 0.40 ~ 1.81mmol/L,超过此范围均视为不正常;(4)糖化血红蛋白:正常值小于等于 6.5%,大于 6.5% 视为不正常;(5)腰臀比:正常值男性小于 0.9,女性小于 0.8;(6)吸烟:包括既往吸烟和正在吸烟的被调查者;(7)糖尿病:自报患者和新诊断患者,即无自报糖尿病史,但本次测定空腹血糖 ≥7.0 mmol/L 者。

3. 统计学分析

使用 SPSS 随机数功能将数据集按 3:1 分为训练数据和预测数据。训练数据用于计算参数和建立模型,预测数据用于评估预测效果。

(1)logistic 回归:模型采用最大似然估计前进法,入选变量和剔除变量的标准分别是是 $P<0.05$ 和 $P>0.10$。

(2)BP 神经网络:采用 SPSS 17.0 统计软件中的神经网络模块的多层感知器。输入层变量为研究所纳入的 10 个自变量,输出层为是否发生糖尿病,定义隐藏层数为 1。

(3)决策树分析:选择卡方自动交互检测,使用分割样本进行验证,无交叉验证,树深度最大值为 3。

(4)受试者工作特征曲线(receiver operator characteristic curve,ROC 曲线):比较 ROC 曲线下面积(area under curve,AUC),最大者表示预测价值最佳。*AUC* 值为 0.5 时,表明无诊断价值,首先要对 *AUC* 与 0.5 的差异进行统计学检验。*AUC* 越接近 1,价值越大。不同模型 *AUC* 的比较用统计量为 *Z* 的非参数检验。所有统计分析均由 SPSS 17.0 和 Medcalc 完成,$P<0.05$ 被认为差异具有统计学意义。

* 基金项目:国家自然科学基金(81373100)
△ 通信作者:吴学森,E-mail:xuesenwu@163.com

结　果

1. 一般情况

共调查社区居民 3077 人。糖尿病患者 550 人，占调查对象的 17.8%。按 1∶2 的原则在与病例生活在相同社区及工作性质相近的正常人群中选择对照，即 1100 名非糖尿病患者纳入此次分析中。本研究所选择对象中，女性居民占 57.7%（952 人）略多于男性 42.3%（698 人）。50 岁以上人群占多数，为 69.0%。文化程度普遍偏低，大专及以上人群仅有 158 人，占研究对象的 9.6%。被调查居民中吸烟人群占 29.2%。BMI 和腰臀比不正常者占很大比例，分别为 58.7% 和 77.7%。有 14% 的研究对象有糖尿病家族史。具体信息见表 1。

表 1　调查对象的一般信息

变量	分组	人口数	构成比(%)
年龄	≥50	1139	69.0
	<50	511	31.0
性别	男	698	42.3
	女	952	57.7
文化程度	大专以下	1492	90.4
	大专及以上	158	9.6
糖尿病家族史	有	231	14.0
	无	1419	86.0
BMI	正常	681	41.3
	不正常	969	58.7
腰臀比	正常	364	22.3
	不正常	1270	77.7
吸烟	是	481	29.2
	否	1169	85.3
高血压	是	242	14.7
	否	1408	76.6

2. logistic 多因素分析

将所研究变量纳入分析模型中，共有以下变量具有统计学意义，分别为年龄、*BMI*、糖化血红蛋白、性别、家族史、甘油三酯。结果见表 2。因此建立预测方程：$P = 1/(1 + e(2.799 - 0.845 \times$ 年龄 $- 0.373 \times BMI - 0.885 \times$ 家族史 $- 2.810 \times$ 糖化血红蛋白 $- 0.588 \times$ 性别 $- 0.679 \times$ 甘油三酯$))$。根据所建方程对预测集数据进行预测，其一致率为 80.8%，ROC 曲线下面积及 95% *CI* 为 0.739(0.694 ~ 0.781)。

表 2　糖尿病风险预测 logistic 多因素分析结果

变量	β	$SE(\beta)$	P 值	OR(95% CI)
年龄	0.845	0.178	0.000	2.33(1.64 ~ 3.30)
BMI	0.373	0.165	0.017	1.45(1.07 ~ 1.97)
糖化血红蛋白	2.810	0.131	0.000	16.67(12.91 ~ 21.54)
性别	0.588	0.150	0.007	1.80(1.34 ~ 2.42)
家族史	0.885	0.212	0.000	2.42(1.60 ~ 3.67)
甘油三酯	0.679	0.157	0.000	1.97(1.45 ~ 2.68)
常数项	−2.799	0.211	−	−

3. BP 神经网络分析结果

所选自变量敏感度分析结果表明对糖尿病发生影响较大的前 5 位因素依此是糖化血红蛋白(0.448)、年龄(0.102)、甘油三酯(0.094)、高血压(0.069)和糖尿病家族史(0.059)。预测数据集结果显示，其预测一致率为 84.1%，ROC 曲线下面积及 95% *CI* 为 0.777(0.734 ~ 0.817)。

4. 决策树分析

树的第一层为糖化血红蛋白，说明糖化血红蛋白与糖尿病关联性最强。其余进入变量依次为年龄、甘油三酯、糖尿病家族史和性别。其预测的一致率为 81.1%，ROC 曲线下面积及 95% *CI* 为 0.737(0.692 ~ 0.779)。

5. ROC 曲线面积比较

三种模型的 ROC 曲线下面积与 0.5 均有统计学差异($P < 0.05$)。三种模型的 ROC 曲线下面积两两比较结果见表 3，结果显示 BP 神经网络模型预测的 ROC 曲线下面积与 logistic 模型($Z = 2.847$, $P = 0.0044$)和决策树模型的 ROC 曲线下面积($Z = 3.050$, $P = 0.0023$)的差异有统计学意义。而 logistic 模型和决策树模型的 AUC($Z = 0.306$, $P = 0.7594$)的差异没有统计学意义。三种模型的 ROC 曲线见图 1。

表 3　三种模型曲线下面积两两比较结果

模型	*SE*	*Z* 值	*P* 值
logistic 回归模型 vs BP 神经网络	0.0135	2.847	0.0044
logistic 回归模型 vs 决策树模型	0.0060	0.306	0.7594
BP 神经网络 vs 决策树模型	0.0132	3.050	0.0023

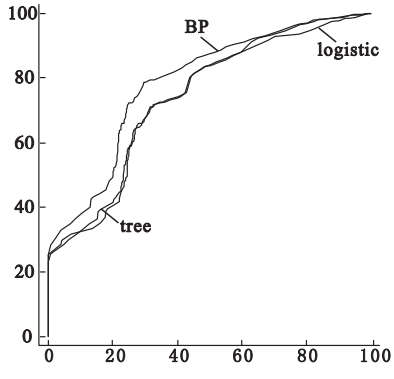

图 1　三种模型的 ROC 曲线

讨　论

1. 糖尿病及其风险预测

糖尿病不仅是威胁人类健康的重要疾病，同时也是很多严重疾病的致病因素，如冠心病、肿瘤等[8]。通过特定的数学模型进行个体糖尿病风险预测，为采取预防干预措施提供建议，有助于提高人群的健康水平和生活质量。本研究通过调查问卷所得变量，建立不同模型进行了比较，显示神经网络模型在预测上具有良好性能。糖尿病是基因与环境共同作用的结果，除了本文所列一些变量外，某些生化标志物如炎性因子、脂联素、microRNA 等也与糖尿病风险有关[9]，但

检测这些成分耗时耗费，并不利于风险评估的快速开展。另外，芬兰等国的糖尿病评分工具，通过一些类似本文的简易的变量都实现出了较好的评价效果[9-10]。

2. BP 神经网络模型

BP 神经网络在医学中有着广泛的应用。徐学琴通过使用 BP 神经网络对全国麻疹的发病率进行了有价值的预测[11]。国外研究分别通过 logistic 回归和神经网络模型预测脑外伤手术术后院内死亡率，神经网络模型表现出明显的优势[12]。BP 神经网络具有很多优点，比如具有较强的非线性映射能力，可以合理提取输入变量和输出变量之间的规则，并进行修改、容错等[4]。但同时 BP 神经网络也存在一定缺陷，比如对于样本量的问题，至今没有明确的公式。关于隐藏层数的设定，多数研究表明，当 BP 神经网络隐藏层数为 1 时，可以达到较好地反映数据规律、特征及获得较好预测效能的作用。本文作者在探讨 BP 神经网络在肺癌并发症预测价值时，比较了不同隐藏层数的预测效果，结果表明隐藏层数为 1 时获得的 ROC 曲线下面积最大[13]，故本研究中 BP 神经网络隐藏层数设定为 1。另外，BP 神经网络无法解释某个变量的作用方向，而 logistic 回归却能对模型和变量具有很好的解释性。

3. 决策树模型及其应用

决策树模型运算时间短，结果以树状显示简单直观，结果的分类把握度较准确。但分类属性增多情况下，会影响预测的效果[14]。决策树模型同 BP 神经网络模型类似，也无法判断某因素的作用方向。以往多数研究显示决策树模型在预测效能上好于 logistic 回归，如决策树在预测高血压患者健康素养中优于 logistic 回归[15]。而本文在糖尿病预测中，两种模型间效果没有统计学差异，可能与树的深度设置、剪接方法有关，需要在以后的研究中进一步探讨。

简洁并快速有效的预测糖尿病风险可以更好地提高全民健康水平。本文研究提示 *BMI* 超标、年龄偏大、男性、糖尿病家族史、糖化血红蛋白均是糖尿病的危险因素。通过数学模型，利用可快速获取的信息进行预测，是未来发展的方向。神经网络模型在预测糖尿病个体风险上有较好的效果。但在实际应用中，logistic 回归对变量有直观的解释，结果容易解释。而神经网络模型和决策树模型对变量却没有很好的解释能力。所以，实际应用中也应结合各自模型的优点，以期在公共卫生实践中取得最好的利用价值。

参 考 文 献

[1] Nathan DM. Diabetes Advances in Diagnosis and Treatment. JAMA, 2015,314(10):1052 - 1062.

[2] Lu C, Sun W. Prevalence of diabetes in Chinese adults. JAMA. 2014, 311(2):199 - 200.

[3] 吴伟，郭军巧，安淑一，等. 使用思维进化算法优化的神经网络建立肾综合征出血热预测模型. 中国卫生统计，2016，33(1)：27 - 31.

[4] 叶华容，杨怡，林萱，等. BP 神经网络在高频彩超特征诊断乳腺癌中的应用. 中国卫生统计，2016，33(1)：71 - 72.

[5] Tseng WT, Chiang WF, Liu SY, et al. The application of data mining techniques to oral cancer prognosis. J Med Syst, 2015, 39(5):59

[6] 高明，唐顺，徐福文. 医院数据挖掘平台中 X - 11 - ARIMA 预测模型的应用研究. 中国卫生统计，2016，33(1)：139 - 141.

[7] Gonzalez GH, Tahsin T, Goodale BC, et al. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. Brief Bioinform, 2016, 17(1):33 - 42.

[8] Leon BM, Maddox TM. Diabetes and cardiovascular disease: Epidemiology, biological mechanisms, treatment recommendations and future research. World J Diabetes, 2015, 6(13):1246 - 1258.

[9] 张晶，金雪娥. 2 型糖尿病患病风险预测的研究进展. 中华实用诊断与治疗杂志，2013，27(9)：839 - 841.

[10] Wannamethee SG, Papacosta O, Whincup PH, et al. The potential for a two - stage diabetes risk algorithm combining non - laboratory - based scores with subsequent routine non - fasting blood tests: results from prospective studies in older men and women. Diabet Med, 2011, 28(1):23 - 30.

[11] 徐学琴，杜进林，孙宁，等. 改进的 BP 神经网络模型在麻疹预测中的应用研究. 中国现代医学杂，2014，24(31)52 - 55.

[12] Shi HY, Hwang SL, Lee KT, et al. In - hospital mortality after traumatic brain injury surgery: a nationwide population - based comparison of mortality predictors used in artificial neural network and logistic regression models. J Neurosurg, 2013, 118(4):746 - 752.

[13] 宋健：logistic 回归模型、神经网络模型和决策树模型在肺癌术后心肺并发症预测中的比较. 安徽医科大学，2014.

[14] 薛允莲. logistic 回归结合决策树技术在冠心病患者住院费用组合分析中的应用. 中国卫生统计，2015，32(6)：988 - 992.

[15] 李现文，李春玉，Kim M，等. 决策树与 logistic 回归在高血压患者健康素养预测中的应用. 护士进修杂志，2012，27(13)：1157 - 1159.

（责任编辑：刘　壮）