doi:10.3772/j.issn.1002-0470.2019.05.006

基于 Logistic 回归和多层神经网络的Ⅱ型糖尿病并发症预测①

王 洁②* 乔艺璇③* 彭 岩* 陈 晓**

(*首都师范大学管理学院 北京 100048)

(**中国人民解放军空军总医院肿瘤放疗科 北京 100142)

摘 要 研究了Ⅱ型糖尿病并发症的预测。针对相关诊断指标众多,直接应用传统的神经网络等模型预测,会带来无法适应多种并发症、运算速度较慢及预测准确率偏低等问题,提出了基于 Logistic 回归和多层神经网络(MNN)的Ⅱ型糖尿病并发症预测模型。该模型首先应用关联性分析,提取与 5 种不同Ⅱ型糖尿病并发症相关的诊断指标,经 Logistic 回归模型等分析得到强相关因子,作为预测模型的输入,再运用 Python,构建基于多层神经网络的预测模型。实验结果表明,全血糖化血红蛋白测定,尿胆原定性实验指标,尿素和尿红细胞与绝大部分Ⅱ型糖尿病并发症直接相关。Logistic 回归结合多层神经网络预测准确率高于单一 Logistic 回归模型,预测准确率基本保持在 85% 的水平上,对某些并发症的预测准确率达到 90% 以上,可以达到为Ⅱ型糖尿病并发症预测提供科学参考的目的。

关键词 Ⅱ型糖尿病并发症,关联因素,多层神经网络(MNN),Logistic 回归,风险预测

0 引 言

糖尿病作为当前医学界已知的并发症最多的疾病,已经严重威胁到人类健康,成为当下最具威胁性的非传染疾病之一[1]。临床数据显示,有 30% ~ 40% 的患者会在糖尿病发病后 10 年左右,至少发生任意一种并发症,且很难通过药物医治将其逆转。Ⅱ型糖尿病是最为常见的糖尿病类型,探究其并发症的影响因素具有重要意义。

近 3 年,针对糖尿病并发症的研究多为关注单一因素的临床试验。糖化血红蛋白、血清胰腺再生蛋白水平证实可用于Ⅱ型糖尿病并发症预测[2,3]。可溶性内皮细胞蛋白 C 受体和高敏 C 反应蛋白水平在糖尿病患者血管并发症的预测中有一定作用[4]。从糖尿病并发症分类来看,研究急性并发症的学者较多,有学者利用 Cox 回归分析预测糖尿病急性并发症继发轻度认知功能障碍[5],也有学者将改良早期预警评分、血糖值评分及两评分结合进行预测[6]。由于研究多由医学类学者进行,因此在研究对象方面,缺乏基于多种因素对多种类并发症的统一研究。

在涉及大量数据分析的研究中,神经网络与 Logistic 回归在疾病预测方面得到广泛应用。包括川崎病并发冠状动脉病变、痢疾、重症手足口病、登革热疫情、麻疹、高血压等[7-12]。文献[13]利用反向传播(back propagation, BP)神经网络基于患者基本信息与生化检查结果,对糖尿病生化指标进行了预测,但是输出结果仅在 3 项血检指标上表现良好,无法表明该模型对某一类型的糖尿病或其并发症的预测意义。文献[14]基于 Logistic 回归探讨了Ⅱ型糖尿病患者糖尿病视网膜病变的危险因素,但只探究

① 国家自然科学青年基金(61601310),北京市教委社科(SM201910028017)和北京市教委科技创新服务能力建设(19530050142,19530050187)资助项目。

② 女,1977 年生,博士,副教授;研究方向:数据挖掘;E-mail:wangjie@cnu.edu.cn

③ 通信作者,E-mail:Qiaoyx97@163.com

(收稿日期:2018-07-25)

了因素与是否患病的相关性,没有应用于预测。文献[15]分析了Ⅱ型糖尿病发生的危险因素,探究了 Logistic 回归、BP 神经网络模型在患病风险预测中的应用。但其模型均用统计产品与服务解决方案(statistical product and service solutions, SPSS)创建,运算速度较慢,且未考虑保护因素对患病与否的影响。

本文基于国家人口与健康科学数据共享服务平台糖尿病患者的数据集,结合 Pearson 相关分析等统计学方法,提出了基于 Logistic 回归和多层神经网络的Ⅱ型糖尿病并发症预测模型。在较大规模真实数据集下,通过与单一 Logistic 回归预测模型的对比实验分析,研究人工神经网络在多种Ⅱ型糖尿病并发症预测方面的应用。

1 Logistic 回归

Logistic 回归(Logistic regressive)是一种广义线性回归,在医学领域被广泛使用,可以运用于寻找疾病的危险因素和预测判别在不同自变量下发生某病的概率。

Logistics 回归模型的基本架构来自多元线性回归模型,即:

$$\hat{P} = \alpha + \beta_1 x_1 + \cdots + \beta_m x_m \tag{1}$$

由于该模型左右两边的取值范围不同,并且反应变量 P 与自变量的关系通常呈 S 型曲线关系,在 1970 年,Cox 引入 Logit 变换。建立以 $logit(P)$ 为因变量拥有 p 个自变量的 Logistic 回归模型如下:

$$logit(P) = \beta_0 + \beta_1 x_1 + \cdots + \beta_P x_P \tag{2}$$

对上面的式子进行逆推可得:

$$P = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_P x_P)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_P x_P)} \tag{3}$$

$$1 - P = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_P x_P)} \tag{4}$$

上述 3 个公式分别为 Logistic 回归模型的不同形式,根据标化的 β 值大小可确定各因素对因变量的影响大小。

本文在预测阶段对患病情况进行重新编码,令患病 = "1",不患病 = "0",并使患病情况作为因变量,各病的关联因素分析后筛选的指标作为因子和协变量。预测集与检验集比例为 7:3。模型的预测结果根据其 Hosmer 和 Lemeshow 检验的显著值判断模型的拟合度,根据预测的准确性判断模型的预测效果。

2 多层神经网络

多层神经网络(multilayer neural network, MNN)是一种对数据进行表征学习的基于机器学习的方法。多层神经网络的优势在于用非监督式或者是半监督式特征学习,以及分层特征提取高效算法来替代手工。与传统训练方式不同,MNN 能使训练多层神经网络的时间大幅度缩短。因此,近些年关于多层神经网络的研究与应用不断出现。

多层神经网络的结构符合经典的神经网络结构,区别在于隐藏层。图 1 为拥有 2 层隐藏层的神经网络模型。

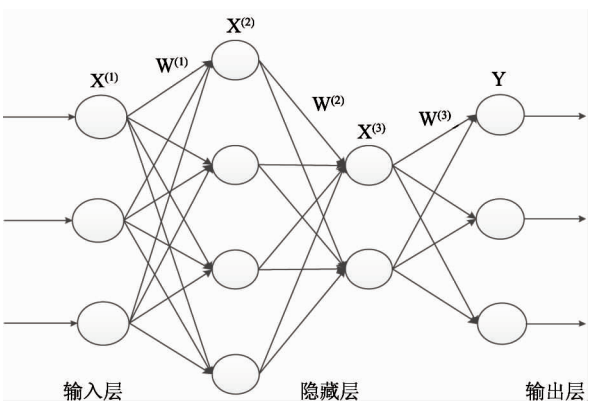

图 1 多层神经网络结构

多层神经网络模型具体包括 2 个过程。第一是前馈过程,输入层的神经元个数 n 代表 n 个相关因子,隐含层将输入层的运算结果作为输入,将隐含层输出作为输出层的输入值,并利用成本函数计算误差值用于调整模型参数。其中,单个隐含层和输出层的神经元输出与输入之间的函数关系为

$$I_j = \sum_i W_{ij} O_i + \theta_j \tag{5}$$

$$O_j = ReLU(I_i) \tag{6}$$

公式中的 W_{ij} 表示神经元 i 与 j 之间连接的权重,O_j 是神经元 j 的输出。为提高模型运算速度,本文选取 $ReLU$ 函数作为激励函数。

第二是反向传播过程,通过对比上一过程得到的预测值与观测值之间的误差调整权重的过程,从而实现网络的训练。

输出层:根据预测结果与真实数据的比较,计算误差。在式(7)中,T_k 表示数据集本身的标签,即通过输入层进入模型的观测值,O_k 表示经各层分析预测得到的预测值。

$$Err(k) = O_k(1 - O_k)(T_K - O_k) \tag{7}$$

隐藏层:

$$Err(j) = O_j(1 - O_j)\sum_K Err_{(k)} \times W_{ij} \tag{8}$$

训练过程中根据隐藏层误差值更新权重:

$$\Delta W_{ij} = (l)Err_j \times O_i \tag{9}$$

$$W_{ij} = W_{ij} + \Delta W_{ij} \tag{10}$$

偏向更新:

$$\Delta\theta_j = (l)Err_j \tag{11}$$

$$\theta_j = \theta_j + \Delta\theta_j \tag{12}$$

本文的模型在训练过程中主要调整 3 个参数:每个隐藏层的神经元数;训练模型需要的迭代次数;输入层影响因素的选择,选取预测效果最佳的 3 个参数值以此来设置整个模型。各种并发症的网络训练的学习率均定为 0.05,隐藏层神经元的个数在训练时随机修改,直至训练误差最小。将各数据库随机按 7:3 比例分成训练组和测试组。随机分配训练数据且多次调整数据输入顺序,从而降低模型的出错率。

3　实验结果与分析

本文首先运用单因素分析对病人各项生理指标进行关联因素分析,再利用分析结果,分别使用单一 Logistic 回归模型以及基于 Logistic 回归和多层神经网络的模型,对Ⅱ型糖尿病并发症进行预测分析。将模型所得结果分别与实际情况进行对比,并阐述 2 种模型各自的优缺点。实验预测结果及对结果的分析说明,基于 Logistic 回归和多层神经网络的模型在Ⅱ型糖尿病并发症预测中具有较高的应用价值。

3.1　实验数据

本文数据来自国家人口与健康科学数据共享服务平台糖尿病患者数据集,包括诊断表、尿常规表、生化表和糖化表等 22 张表格。以诊断表的 patient_id(病人 ID)作为关键字对所有表格进行合并,筛选所有诊断描述字段包含Ⅱ型糖尿病以及其同义词的个体组成新的数据库,共有个案 7 499 例。

利用数字 1 ~8 对诊断描述变量重新编码,分别对应Ⅱ型糖尿病及其同义词、Ⅱ型糖尿病性心脑血管并发症、Ⅱ型糖尿病性眼部并发症、Ⅱ型糖尿病性神经病变、Ⅱ型糖尿病性伴有多个并发症、Ⅱ型糖尿病性肾病、Ⅱ型糖尿病性急性并发症及Ⅱ型糖尿病性足病。7 种Ⅱ型糖尿病并发症患者数据分别与未患并发症的Ⅱ型糖尿病患者数据融合,形成对应数据库。

医学检测数据采用不同方法进行变量处理。原数据库的部分指标变量值中包含中文文字“乳糜”、“溶血”、“已复查”等异常值,去除中文并转换成数值型。对于变量“铁”、“年龄”、“性别”、“脂肪酶”、“尿比重测定”、“尿液颜色”等缺失值超过一半的变量,查阅相关文献后认为与Ⅱ型糖尿病并发症发病无显著相关,直接舍去,以免其影响整体结果。经过多种缺失值填充方法比较,其余变量采取期望最大化算法(expectation maximization, EM)进行缺失值填充。通过后文的模型分析结果,可以验证该方法在糖尿病数据集中适用性较高。对于每一个指标变量,根据原始数据给出的指标正常范围并查阅相关资料,按照关联因素分析的要求进行重新量化。各变量对应代表符号见表 1,X_i 代表自变量,P_i 代表因变量。

表 1　变量对照表

变量符号	变量名称	变量符号	变量名称
X_1	全血糖化血红蛋白	X_{14}	γ-谷氨酰基转移酶
X_2	丙氨酸氨基转移酶	X_{15}	尿蛋白定型试验
X_3	尿素	X_{16}	尿浊度
X_4	钾	X_{17}	总胆固醇
X_5	肌酐	X_{18}	无机磷
X_6	葡萄糖	X_{19}	镁
X_7	尿白细胞	P_1	Ⅱ型糖尿病心脑血管并发症

续表 1

X_8	尿胆原定性试验	P_2	Ⅱ型糖尿病性眼部并发症
X_9	尿红细胞	P_3	Ⅱ型糖尿病性神经病变
X_{10}	尿酵母细胞数	P_4	Ⅱ型糖尿病伴有多个并发症
X_{11}	尿酮体试验	P_5	Ⅱ型糖尿病性肾病
X_{12}	尿浊度	P_6	Ⅱ型糖尿病急性并发症
X_{13}	血清白蛋白	P_7	Ⅱ型糖尿病性足病

3.2 单一 Logistic 回归预测分析结果

本文在 SPSS20.0 环境下进行相关设置,首先依据 Person 相关系数进行单因素分析,根据分析结果,利用 Logistic 回归模型进行预测,预测结果见表 2。

表 2 Logistic 回归模型训练结果

名称	显著性值	准确率(%)
P_1	0.347	82.7
P_2	0.000	68.0
P_3	0.019	71.9
P_4	0.274	77.9
P_5	0.440	71.2
P_6	0.004	15.3
P_7	0.007	27.7

对Ⅱ型糖尿病性各并发症的预测,Logistic 回归的预测大多低于 80%,准确率整体偏低。从 Hosmer 和 Lemeshow 检验的显著性值来看,有 4 个并发症预测模型的显著性值 <0.05,这表示该并发症预测模型拟合度不够,模型预测与现实情况差距较大,该 2 个模型应用的实际意义不够理想;其他 3 个模型的显著性值 >0.05,这表示该并发症预测模型拟合度较好,模型预测能够较真实模拟该并发症的患病情况。

其中,对Ⅱ型糖尿病心脑血管并发症的预测,Logistic 回归的预测准确率达到 82.3%,该模型的拟合度和准确性都较高,可供参考。对应的接受者操作特征曲线(receiver operating characteristic,ROC)如图 2 所示;对Ⅱ型糖尿病性眼部并发症、Ⅱ型糖尿病性神经病变、Ⅱ型糖尿病急性并发症和Ⅱ型糖尿病性足病的预测,模型的拟合度和准确性都不够理想,无法准确预测这四个并发症的患病情况;对Ⅱ型糖尿病心脑血管并发症、Ⅱ型糖尿病伴有多个并发症和Ⅱ型糖尿病性肾病的预测,模型的拟合度较好但准确性不够理想。基于以上分析,尝试将 Logistic 回归多因素分析结果与其他模型相结合,以期获得更准确的预测结果。

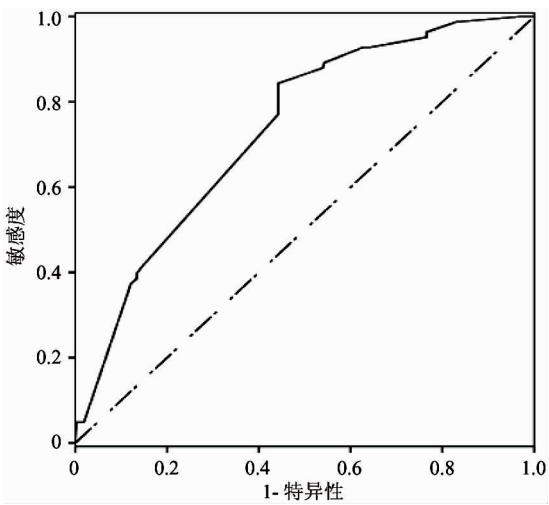

图 2 Ⅱ型糖尿病心脑血管并发症预测 ROC 曲线

3.3 Logistic 回归和多层神经网络预测模型分析结果

根据 3.2 节中的单因素分析结果,本节利用多因素 Logistic 回归模型对各个并发症进行直接关联因素分析,分析所得强相关因子作为多层神经网络预测模型的输入。

3.3.1 Logistic 回归获取强相关因子

Logistic 回归分析得到的结果根据 p 值确定各因素是否与该并发症相关($p<0.05$ 时具有统计学意义),根据 β 值的正负判断各关联因素是否为危险因素,$\beta>0$ 时则为危险因素,标化的 β 值确定各因素对因变量的影响大小。各个并发症发病关联因素分析结果见表 3。

由于最终得出仅有 2 个因素都与Ⅱ型糖尿病急性并发症显著相关。可供输入的关联因素过少,会导致预测模型的预测结果误差过大,故不展示Ⅱ型糖尿病急性并发症的模型预测结果。数据集中Ⅱ型

表 3　关联因素分析结果

并发症	保护因素	危险因素
P_1	X_4, X_9	X_1, X_3, X_8
P_2	$X_2, X_5, X_6, X_9, X_{10}, X_{11}$	$X_1, X_3, X_8, X_7, X_{12}$
P_3	$X_2, X_4, X_9, X_{10}, X_{15}$	$X_1, X_7, X_8, X_{12}, X_{13}, X_{14}$
P_4	X_2, X_9, X_{17}, X_{18}	$X_1, X_3, X_8, X_{12}, X_{14}$
P_5	X_2, X_9, X_{10}	$X_1, X_3, X_8, X_{14}, X_{15}, X_{16}, X_{19}$
P_6	X_9	X_{11}

糖尿病性足病的有效数据过少,造成关联因素分析的误差过大,最终得出所有因素都与Ⅱ型糖尿病性足病无显著相关,故不展示对于糖尿病性足病关联分析及模型预测结果。

由各并发症发病关联因素 Logistic 回归表达式的 β 值得到尿红细胞、丙氨酸氨基转移酶、尿酵母细胞等指标值是大多数Ⅱ型糖尿病并发症的保护因素,全糖化血红蛋白、尿素、尿胆原定性试验、尿白细胞和尿浊度则是大多数Ⅱ型糖尿病并发症的危险因素,对于Ⅱ型糖尿病患者,重点监测这几项指标对其并发症的预防和治疗具有积极意义。

3.3.2　多层神经网络预测结果

模型采用 Python 语言编写,经过多次修改训练参数及调整实验数据,最终得出训练模型的最优结果。具体结果见表 4,误差为反向训练时误差达到阈值时的最终数值;准确率为当误差在 0.05 的范围内时对应的比例。准确率 = 预测正确的结果总数/总患者人数。

表 4　模型训练结果

名称	P_1	P_2	P_3	P_4	P_5
误差	2.52	2.707	0.048	2.78	5.77
准确率(%)	89.86	88.3	92.5	87.67	83.33

整体来说,Logistic 回归加多层神经网络对Ⅱ型糖尿病并发症的预测准确率保持在 85% 的水平上,对某些并发症的预测准确率可以达到 90% 以上,其中Ⅱ型糖尿病性神经病变的预测中,由于其相关因素数目多,预测结果也更准确。表中Ⅱ型糖尿病性肾病和Ⅱ型糖尿病伴有多个并发症等疾病的预测率偏低,可能与相关因素的选择有关,推测相关因素中可能存在关联性比较弱的因素,可以再进行进一步的测试。

预测准确率受多重因素的影响,有相关因素的个数(即输入层神经元数)、隐藏层神经元数,输入数据的顺序、数目等因素。总体来说,该模型对Ⅱ型糖尿病并发症的预测具有较高的准确率,对其发病的预测具有实际意义。

3.4　模型对比

2 个模型对 5 种并发症预测的准确率对比如图 3。图中纵轴表示各模型预测的准确率,横轴对应表 1 中的 5 种并发症。斜线柱状图表示单一 Logistic 回归模型,灰色柱状图表示 Logistic 回归和多层神经网络模型。

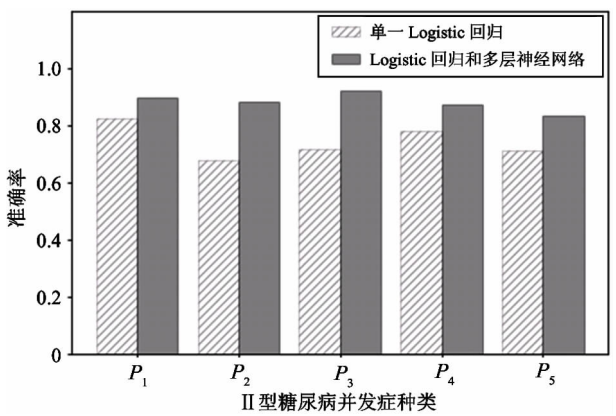

图 3　2 种模型准确率对比

对比发现,单一 Logistic 回归对于Ⅱ型糖尿病心脑血管并发症的预测准确率较其他并发症高,但对于其他并发症,单一 Logistic 预测模型准确率过低,模型拟合度也不够理想,预测效果不佳。

相较之下,结合 Logistic 回归和多层神经网络构建的预测模型经过训练后达到的预测结果准确率基本在 85% 以上,对某些特定的疾病预测准确率甚至可达 90%,预测效果更为理想,说明该方法在Ⅱ型糖尿病并发症预测上具有较高的应用价值。

4　结　论

本文根据关联因素分析结果可知,全血糖化血

红蛋白测定,尿胆原定性试验指标,尿素和尿红细胞与绝大部分Ⅱ型糖尿病并发症直接相关。故在临床检查中,要着重对Ⅱ型糖尿病患者的以上 4 个指标进行监测,在记录电子病历的过程中,可将以上 4 个指标取出成立数据库用于观测,并注意保存数据,进行波动性分析。根据分析结果,患者的全血糖化血红蛋白指标要重点监测,对于该指标有异常的患者,要加强对其并发症发病的预防与检查。

在真实数据集上进行的对比实验发现,与单一 Logistic 回归模型相比,基于 Logistic 回归和多层神经网络的模型在多种Ⅱ型糖尿病并发症预测中具有较高的准确率,对特定并发症的预测准确率可以达到 90%。实验证明了本文方法的可行性,该方法可为Ⅱ型糖尿病并发症的预防、诊断和治疗提供一定的参考。

如何筛选一个病人的多条数据是数据处理过程中的关键问题,本文对于微小波动的数据,选均值为代表;对差距较大的数据,根据病人确诊时间及波动情况选取代表值。针对研究过程中数据缺失、部分表示内涵不明的问题,应与医疗机构进行沟通,注意数据填写的规范性,便于数据分析工作者理解数据。因此,后续研究将继续收集相关数据,扩充数据的来源与形式,增强预测结果的可信度。

参考文献

[1] 中华医学会糖尿病学分会. 中国 2 型糖尿病防治指南(2017 年版)[J]. 中国实用内科杂志, 2018, 38(4): 292-344

[2] 吕福应,郭国才,黎敦镇, 等. 糖化血红蛋白对糖尿病诊疗和并发症风险预测的临床价值[J]. 国际检验医学杂志,2015,36(11):1591-1592

[3] 李玲,杨家悦,祝祥云, 等. 血清胰腺再生蛋白水平预测 2 型糖尿病及其慢性并发症发生的价值[J]. 中国全科医学,2016,19(2):159-163

[4] 蒙绪标,符兰芳,刘婷婷, 等. 可溶性内皮细胞蛋白 C 受体和高敏 C 反应蛋白水平预测糖尿病患者血管并发症的价值[J]. 中国现代医学杂志, 2016,26(14): 58-62

[5] 明淑萍,刘玲,周黎, 等. 糖尿病急性并发症继发轻度认知功能障碍的预测模型及时间窗分析[J]. 中风与神经疾病杂志,2017,34(9):786-791

[6] 李晓燕,孟凡杰,段玉龙,等. 改良早期预警评分、血糖值评分及两评分结合预测糖尿病急性并发症患者预后能力的对比研究[J]. 实用医学杂志, 2018,34(3): 397-400

[7] 张胜,田杰,樊楚, 等. 基于神经网络的川崎病并发冠状动脉病变预测模型[J]. 中国生物医学工程学报, 2018,37(3):313-318

[8] 肖达勇,刘勋,廖骏, 等. 2009-2014 年重庆市痢疾流行特征及气象因素对其影响的 BP 神经网络模型研究[J]. 预防医学情报杂志,2018,34(6):722-727

[9] 王斌,冯慧芬,黄平,等. 人工神经网络模型在预测重症手足口病中的应用研究[J]. 现代预防医学,2018, 45(11):1921-1924 + 1947

[10] 任红艳,吴伟,李乔玄,等. 基于反向传播神经网络模型的广东省登革热疫情预测研究[J]. 中国媒介生物学及控制杂志,2018,29(3):221-225

[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[J]. 中国现代医学杂志, 2014,24(31):52-55

[12] 何朝,胡建功,赵莹颖, 等. 北京市社区高血压患者血压控制影响因素 Logistic 回归分析[J]. 现代预防医学,2018,45(7):1211-1215

[13] 陈德华,洪灵涛,潘乔. 基于改进神经网络的糖尿病生化指标值预测[J]. 微型机与应用,2017,36(5):54-56 + 59

[14] 袁媛. 2 型糖尿病患者糖尿病视网膜病变的预测性因素研究[J]. 眼科新进展, 2015,35(8):784-786

[15] 陈渝,宗会娟,李伟. 2 型糖尿病危险因素及患病风险预测模型研究[J]. 昆明理工大学学报(自然科学版), 2018,43(2):60-64 + 70

Prediction of type Ⅱ diabetes complications based on logistic regression and multilayer neural network

Wang Jie*, Qiao Yixuan*, Peng Yan*, Chen Xiao**

(*School of Management, Capital Normal University, Beijing 100048)

(**Department of Radiation Oncology, Air Force General Hospital, Beijing 100142)

Abstract

The prediction of type Ⅱ diabetes complications is studied. The diagnostic criteria of type Ⅱ diabetes complications are numerous, and direct application of traditional neural network model prediction will bring about the situation that it cannot adapt to many kinds of complications, slow operation speed and low prediction accuracy. A prediction model for type Ⅱ diabetes complications based on Logistic regression and multilayer neural network (MNN) is proposed. The model firstly uses correlation analysis to extract diagnostic indexes related to five different types of type Ⅱ diabetic complications, and the strong correlation factor is obtained through Logistic regression model analysis, and then a prediction model based on multilayer neural network is constructed by using Python language. The experiment results show that the measurement of total glycemic hemoglobin, urinary biliary original test indicators, urea and urinary red blood cells are related to most of the complications of type Ⅱ diabetes directly. The prediction accuracy of multi-layer neural network is higher than that of Logistic regression model. The prediction accuracies are maintained at 85% basically, and the prediction accuracy of certain complications can reach more than 90%, which provides a scientific reference for the complications of type Ⅱ diabetes.

Key words: type Ⅱ diabetes mellitus complications, associated factors, multilayer neural network (MNN), Logistic regression, risk prediction