DOI:10.3969/j.issn.1005-202X.2018.10.022

医学生物物理

基于K-Nearest Neighbor和神经网络的糖尿病分类研究

陈真诚[1],杜莹[2],邹春林[3],梁永波[1],吴植强[4],朱健铭[1]
1.桂林电子科技大学生命与环境科学学院,广西 桂林 541004;2.桂林电子科技大学电子工程与自动化学院,广西 桂林 541004;
3.广西医科大学转化医学研究中心,广西 南宁 530021;4.广西壮族自治区医疗器械检测中心,广西 南宁 530021

【摘 要】为实现糖尿病的早期筛查,提高对糖尿病分类的准确度,在研究有关糖尿病危险因素的基础上,增加糖化血红蛋白作为糖尿病早期筛查的特征之一。研究中选取与人类最为相似的食蟹猴作为研究对象,利用年龄、血压、腹围、BMI、糖化血红蛋白以及空腹血糖作为特征输入,将正常、糖尿病前期和糖尿病作为类别输出,利用K-Nearest Neighbor(KNN)和神经网络两种方法对其分类。发现在增加糖化血红蛋白作为分类特征之一时,KNN(*K*=3)和神经网络的分类准确率分别为81.8%和92.6%,明显高于没有这一特征时的准确率(68.1%和89.7%),KNN和神经网络都可以对食蟹猴数据进行分类和识别,起到早期筛查作用。

【关键词】糖尿病;糖化血红蛋白;空腹血糖;KNN;神经网络;食蟹猴
【中图分类号】R318;Q819 【文献标志码】A 【文章编号】1005-202X(2018)10-1220-05

Classification of diabetes based on K-Nearest Neighbor and neural network

CHEN Zhencheng[1], DU Ying[2], ZOU Chunlin[3], LIANG Yongbo[1], WU Zhiqiang[4], ZHU Jianming[1]
1. School of Life and Environmental Sciences, Guilin University of Electronic Technology, Guilin 541004, China; 2. School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin 541004, China; 3. Transforming Medical Research Center, Guangxi Medical University, Nanning 530021, China; 4. Medical Devices Testing Center of Guangxi Zhuang Autonomous Region, Nanning 530021, China

Abstract: In order to achieve the early screening for diabetes and improve the accuracy of classification of diabetes, on the basis of studying the risk factors of diabetes, glycosylated hemoglobin is added as one of the features in the early screening for diabetes. Herein cynomolgus monkeys that are most similar to humans were selected as the study subjects. Several features, such as age, blood pressure, abdominal circumference, body mass index, glycated hemoglobin and fasting blood glucose, are chosen as inputs, while normal, prediabetes and diabetes are output as categories. Both K-Nearest Neighbor (KNN) and neural network are used to classify diabetes. When glycosylated hemoglobin is added as one of the classification features, the accuracy of classification using KNN (*K*=3) and neural network is 81.8% and 92.6%, respectively, significantly higher than the accuracy which is obtained without considering the feature of glycosylated hemoglobin (68.1% and 89.7%). Therefore, both KNN and neural network can classify and identify the data of cynomolgus monkey and achieve an early screening for diabetes.

Keywords: diabetes; glycosylated hemoglobin; fasting blood glucose; K-Nearest Neighbor; neural network; cynomolgus monkey

前 言

糖尿病是一种常见的代谢性慢性病[1],随着国民生活水平的提高、生活节奏加快,糖尿病患者的数量逐年上升,并呈低龄化趋势,已成为我国重大的公共健康问题[2-5]。然而绝大多数的患者意识不到糖尿病患病前期出现的症状,以致于最终发展为糖尿病[6-7]。糖耐量是机体对葡萄糖的耐受能力,若在发病前的潜伏期能够积极采取相应的干预措施,每年大约有6%~10%[8]的患者将不会发展为糖尿病。国内外有研究指出,相比于糖耐量正常(Normal Glucose Tolerance, NGT),糖耐量受损(Impaired Glucose

【收稿日期】2018-05-25
【基金项目】国家自然科学基金重大科研仪器研制项目(61627807);广西自然科学基金(2017GXNSFGA198005);国家重点研发计划课题(2016YFC1305703);广西自然科学基金青年基金(2016GXNSFBA380145);广西自动检测技术与仪器重点实验室主任基金(YQ17118);广西信息科学实验中心一般项目(YB1513)
【作者简介】陈真诚,教授,博士生导师,研究方向:生物传感与智能仪器,E-mail: 18078842451@163.com;杜莹,研究生,研究方向:生物传感与智能仪器,E-mail: 18946072216@163.com
【通信作者】吴植强,副主任技师,主要从事医疗器械检验检测和管理工作,E-mail: wzqnn@126.com;朱健铭,博士,副教授,硕士生导师,研究方向:生物传感与智能仪器,生物医学信号处理,E-mail: zjmcsu@126.com

Tolerance, IGT)人群更容易发生其他心血管疾病[9]，糖尿病及其并发症已对人类健康造成严重的威胁，于是对糖耐量进行有效评估显得尤为重要，如果可以在患病早期做出及时的诊断和治疗[10]，就可以预防糖尿病。在已报导的动物模型中，非人灵长类动物的糖尿病病程、病症和人类的糖尿病极为相似[11]，本文选取非人灵长类中的食蟹猴作为研究对象。在以往的糖尿病诊断中，临床上常用的检测手段是口服糖耐量实验(Oral Glucose Tolerance Test, OGTT)，但实验过于繁琐，需多时间点采血。有学者提出建立简单模型对糖尿病进行筛查[12-13]，本文在有关糖尿病高危因素的基础上结合糖化血红蛋白，利用KNN(K-Nearest Neighbor)和神经网络两种方法对食蟹猴进行分类，为日后辅助筛查糖尿病人群做准备。

1 KNN与神经网络模型

KNN分类算法是数据挖掘分类技术中最简单的方法之一。KNN算法的核心思想是如果一个样本在特征空间中的K个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性[14]。该方法在确定分类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

人工神经网络(Artificial Neural Networks, ANN)是由大量的、简单的处理单元(称为神经元)广泛地互相连接而形成的复杂网络系统，它反映了人脑功能的许多基本特征，是一个高度复杂的非线性动力学习系统。单个神经元的结构图如图1所示[15-16]。

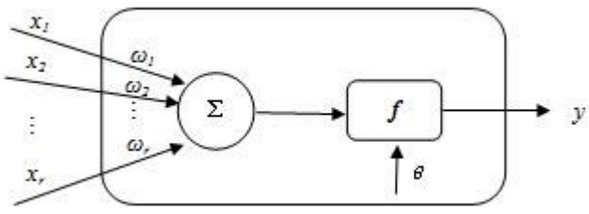

图1 单个神经元结构图

Fig.1 Single neuron structure

在神经元结构中，$\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_r$是神经元的输入向量，$\boldsymbol{y}$是神经元的输出向量，$\omega_1,\omega_2...\omega_r$是各神经元之间的连接权值，$f$是激活函数(传递函数)，神经元模型中还包括一个外部偏置θ，适当增加或减少激活函数的网络输出。可以用以下公式描述一个神经元:

$$\boldsymbol{y}=f(\sum_{i=1}^{r}\boldsymbol{x}_i\omega_i+\theta)$$

神经网络一般包含一个输入层，一个或多个隐含层以及一个输出层，隐含层是整个网络的核心，与神经网络的性能密切相关。模式识别就是机器识别或计算机识别，其目的在于让机器自动识别事物，使机器具备人所具有的对各种事物与现象进行分析、描述与判断的部分能力。研究目的就是利用计算机对物理对象进行分类，在错误概率最小的条件下，使识别的目的尽量与客观事物相符合。一个典型的神经网络模式识别系统如图2所示[17]。

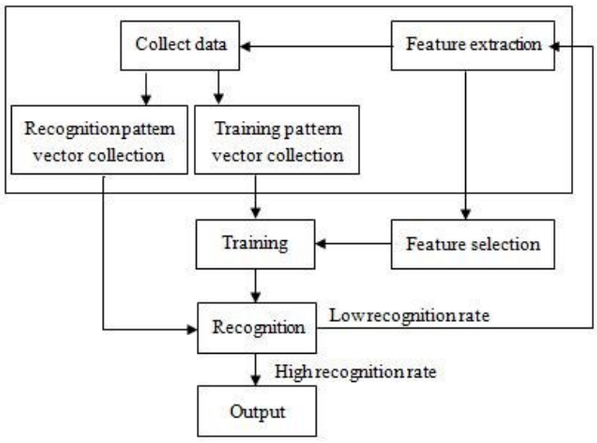

图2 神经网络模式识别过程流程

Fig.2 Neural network pattern recognition process

总的来讲，采集系统将采集到的数据输入模式识别系统，这些数据会形成一个激励向量，寻找激励间的相关属性，这是系统最基本的要求。模式识别的具体过程大致是对研究对象进行数据采集、数据预处理、特征提取和选择以及模式分类4步。

2 实验过程

2.1 数据来源

本文数据采集于南宁灵康赛诺科生物科技有限公司，经6周的试验获取90组成年食蟹猴实验数据。实验开始前一个月对食蟹猴进行猴椅保定、模拟灌胃、模拟采血以及适应性训练，尽量减小实验过程中由于过度紧张而对实验数据造成影响。正式实验前一天，早晨空腹称体质量，下午训练猴子的同时测量血压，要求在动物平静状态下使用美国GE公司Dash 2500监护仪多次测量，取平均值。在禁食至少8 h后，空腹状态下采集血样，按照4 g/(kg·8 mL)配置糖溶液，灌胃行OGTT实验。血样采集后在冰盒内静置30 min，使用湘仪H-2050R离心机进行离心，对血清分装处理，使用罗氏cobas c311全自动生化分析仪检测血清血糖和糖化血红蛋白含量。

2.2 数据处理

在温度16~26 ℃、湿度40%~70%的动物操作间，动物平静状态下采集血样。由于糖尿病食蟹猴模型与人类最为相似，于是在选取糖尿病食蟹猴发病因素时参考了以往人类糖尿病模型，总结对糖尿病影响比较大的高危因素，比如年龄、身高、体质量等。有相关研究表明，糖化血红蛋白可以作为糖尿病筛选的辅助性指标[18-19]，于是将其纳入特征输入行列，试验中对这一指标能否在糖尿病的分类算法中起作用将进行验证。

目前，国内外对动物糖尿病诊断标准尚无统一规定，多数除了观察实验动物“三多一少”症状外，根据文献或设置空白对照组进行比较，有用空腹血糖≥6.1 mmol/L，和（或）随机血糖或糖耐量实验后2 h血糖值≥11.1 mmol/L作为评定标准[20]，还有根据空腹血糖高于5 mmol/L作为患病组和正常组的分界线[21]，或者利用不同体重指数评估糖耐量[22]，结合其他研究人员的工作，根据空腹血糖、糖耐量实验后2 h血糖值、BMI以及糖耐量曲线对食蟹猴数据进行划分。划分类别为：正常猴、糖耐量受损猴和糖尿病猴。划分后食蟹猴基本信息如表1所示。

由于特征的计量单位不同，会导致在训练过程中变动值大的特征权重越来越大，而变动值小的特征被网络认为无关特征。因此，模型训练之前必须使用归一化的方法使所有数据的变动范围一致。选取年龄、舒张压、收缩压、腹围、BMI、皮脂厚度、糖化血红蛋白以及空腹血糖值作为特征输入，输出变量为DM/IGT/NGT这3种状态。分别利用KNN和神经网络模式识别对这3种类别进行识别，在选取糖尿病猴模型中起到辅助作用，进而可以应用在糖尿病及糖尿病前期人群的辅助筛查。

表1 食蟹猴基本信息表（$\bar{x}\pm s$）

Tab.1 General information of cynomolgus monkey (*Mean*±*SD*)

Parameter	Normal group	Prediabetes group	Diabetic group
Age/year	16.00±9.00	17.00±9.00	22.00±4.00
Systolic pressure/mmHg	122.50±20.50	126.50±27.50	119.50±27.50
Diastolic pressure/ mmHg	58.00±15.00	64.50±13.50	64.00±22.00
Abdominal circumference/cm	43.00±16.00	43.00±16.00	34.50±12.50
BMI/kg·m^{-2}	46.56±17.26	40.56±12.76	33.33±4.73
Sebum thickness/mm	11.75±6.95	11.77±6.97	7.58±5.38
GHb/%	3.85±0.35	4.45±0.65	8.40±3.80
FBG/mmol·L^{-1}	4.11±1.37	5.79±2.24	13.84±7.48

BMI: Body mass index;GHb: Glycated hemoglobin; FBG: Fasting blood glucose

3 结果分析

对实验过程采集到的90组数据进行训练数据和测试数据划分，其中68组数据用作训练数据，剩余22组数据用于测试。首先用Matlab对食蟹猴数据使用KNN进行分类研究，为了验证糖化血红蛋白是否可以作为判断糖尿病类型的特征输入，分别使用8个特征（糖化血红蛋白作为特征输入之一）和7个特征（糖化血红蛋白不作为特征输入之一）这两种情况进行说明，对于不同的K值，分类效果有所不同，从表2结果来看，对于使用糖化血红蛋白这一特征的情况下，K值越大，算法的识别准确率越低。而在未使用糖化血红蛋白这一特征时，K=3时的准确率明显低于使用这一特征时的准确率，但K取其它值时是否使用糖化血红蛋白特征对准确率并没有影响。

表2 不同K值对准确率的影响（%）

Tab.2 Effect of different *K* values on accuracy (%)

Number of feature inputs	*K*			
	3	5	7	9
8	81.8	77.2	68.1	63.6
7	68.1	77.2	68.1	63.6

与KNN对比，使用Matlab创建网络，使用均方误差性能函数（Mean Squared Error, MSE）评估网络性能，训练集中的数据依次输入模型，训练算法根据该数据集在模型中的误差不断修正各个节点的权重和偏置。对网络采用反向传播算法验证用于测量模

型的泛化性能，当模型的泛化性能停止增长时则停止训练。测试集对训练过程没有影响，用于独立的评估训练中及训练后的模型性能。

在神经网络中也使用同样的方法对比使用8个特征和7个特征的差别，使用trainscg这一训练函数得到结果如表3所示。从结果可以看到，使用8个特征输入得到的效果要好于7个特征输入。这一结果再一次说明增加糖化血红蛋白作为输入特征，无论是KNN还是神经网络都可以提高模型的准确性。

由于影响检测结果输出的因素有很多，这里讨论使用不同的训练函数对网络输出的影响。通过计算得到结果如表4所示。

由表4可知，使用不同的训练函数网络输出结果不同，但没有明显的规律，在选取训练函数时要多次尝试，选取最优、最适合的函数即可。通过实验结果可以看到将糖化血红蛋白作为判别糖尿病的特征之一，无论是KNN还是神经网络对糖尿病及其前期都有较好的判别效果，在两者对比之下，神经网络对于食蟹猴数据分类有更好的效果。

表3 不同特征输入对结果的影响(%)

Tab.3 Effect of different feature inputs on the results (%)

Number of feature inputs	Training set	Test set
8	92.8	90.9
7	89.7	81.8

表4 不同训练函数训练集与测试集准确率(%)

Tab.4 Accuracy of different training functions for training set and test set (%)

Set	Trainlm	Trainb	Traingdx	Traingdm	Traingda	Trainscg
Training	92.6	91.2	92.6	89.7	92.6	92.6
Test	90.9	90.9	90.9	81.8	86.4	90.9

4 结 论

研究表明，增加糖化血红蛋白作为特征输入之一可以提高模型的准确率，无论是KNN算法还是神经网络都可以应用于糖尿病食蟹猴的早期筛查中，神经网络的准确率略高于KNN方法。这种利用简单特征对糖尿病进行筛查的方法不但操作简单方便，可以减少采血次数，而且也可以对糖尿病的早期筛查起到很好的辅助作用。不单是可以应用于糖尿病食蟹猴模型的初期筛选上，也可以将其应用于糖尿病人群的早期筛查和识别中，可以作为医疗检验的一种辅助手段。

本次实验由于时间和人力资源较少，采集到的数据覆盖并不全面，后期可增加数据量和覆盖面，可将糖尿病前期阶段分为空腹血糖受损和糖耐量受损。若实验条件允许可适当增加检测指标，例如果糖胺，进一步完善预测模型，对于准确识别糖尿病会有更好的帮助。

【参考文献】

[1] 花琦琦, 刘新凤, 焦青, 等. 基于SQL Server的糖尿病信息管理与分析系统[J]. 中国医学物理学杂志, 2016, 33(11): 1183-1188.
HUA Q Q, LIU X F, JIAO Q, et al. Diabetes information management and analysis system based on SQL Server[J]. Chinese Journal of Medical Physics, 2016, 33(11): 1183-1188.

[2] ASSOCIATION A D. Classification and diagnosis of diabetes[J]. Diabetes Care, 2017, 40(Suppl 1): S11.

[3] SMIT A J, SMIT J M, BOTTERBLOM G J, et al. Skin autofluorescence based decision tree in detection of impaired glucose tolerance and diabetes[J]. PLoS One, 2013, 8(6): e65592.

[4] 白杰. 基于高斯混合模型的糖尿病检测[J]. 电脑知识与技术, 2017, 13(11): 1-2.
BAI J. Diabetes detection based on Gaussian mixture model[J]. Computer Knowledge and Technology, 2017, 13(11): 1-2.

[5] 李飞, 王贻坤, 朱灵, 等. 基于神经网络模式识别的糖尿病无创风险评估方法研究[J]. 光谱学与光谱分析, 2014, 34(5): 1327-1331.
LI F, WANG Y K, ZHU L, et al. Non-invasive risk assessment of diabetes mellitus based on neural network pattern recognition[J]. Spectroscopy and Spectral Analysis, 2014, 34(5): 1327-1331.

[6] 王卫庆. 从最新流行病学数据谈糖尿病前期干预的重要性[J]. 药品评价, 2014(13): 18-21.
WANG W Q. The importance of pre-diabetes intervention from the latest epidemiological data[J]. Drug Evaluation, 2014(13): 18-21.

[7] 潘娟, 张杰文, 谢政权, 等. 中国糖尿病教育模式的研究现状[J]. 现代医院, 2016, 16(8): 1205-1207.
PAN J, ZHANG J W, XIE Z Q, et al. Research status quo of diabetes health education mode in China[J]. Modern Hospital, 2016, 16(8): 1205-1207.

[8] GANG L. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction[J]. Health Inf Sci Syst, 2016, 4(1): 1-9.

[9] BARDINI G, DICEMBRINI I, CRESCI B, et al. Inflammation markers and metabolic characteristics of subjects with 1-h plasma glucose levels [J]. Diabetes Care, 2010, 33(2): 411-413.

[10] 侯玉梅, 朱亚楠, 尹福在. 基于支持向量机和人工神经网络的2型糖尿病患病风险预测研究[J]. 现代预防医学, 2017, 44(11): 1921-1924.
HOU Y M, ZHU Y N, YIN F Z. Predictive risk prediction of type 2

diabetes based on support vector machine and artificial neural network [J]. Modern Preventive Medicine, 2017, 44(11): 1921-1924.

[11] 韦祝梅, 杨波, 李振明, 等. 肥胖及糖尿病食蟹猴全天血糖、胰岛素值及相关生理指标观测[J]. 实验动物与比较医学, 2015, 35(5): 394-397.
WEI Z M, YANG B, LI Z M, et al. Obesity and diabetic cynomolgus monkey whole-day blood glucose, insulin values and related physiological indices [J]. Laboratory Animal and Comparative Medicine, 2015, 35(5): 394-397.

[12] 苏萍, 杨亚超, 杨洋, 等. 健康管理人群2型糖尿病发病风险预测模型[J]. 山东大学学报(医学版), 2017, 55(6): 82-86.
SU P, YANG Y C, YANG Y, et al. Risk prediction model of type 2 diabetes in health management population[J]. Journal of Shandong University (Medical Science), 2017, 55(6): 82-86.

[13] WANG C, LI L, WANG L, et al. Evaluating the risk of type 2 diabetes mellitus using artificial neural network: an effective classification approach[J]. Diabetes Res Clin Pract, 2013, 100(1): 111-118.

[14] SU Y, GUAN S U, HONG W C. Density and distance based KNN approach to classification[J]. Int J Appl Evol Comput, 2016, 7(2): 45-60.

[15] 韩敏. 人工神经网络基础[M]. 大连: 大连理工大学出版社, 2014: 29-30.
HAN M. Foundation of artificial neural network[M]. Dalian: Dalian University of Technology Press, 2014: 29-30.

[16] 刘彦杰. 优化BP神经网络在糖尿病患病风险分析中的应用[D]. 兰州: 兰州理工大学, 2016: 11-12.
LIU Y J. Application of optimized BP neural network in risk analysis of diabetes mellitus[D]. Lanzhou: Lanzhou University of Technology, 2016: 11-12.

[17] 王晓梅. 神经网络导论[M]. 北京: 科学出版社, 2017: 256-257.
WANG X M. Introduction to neural networks[M]. Beijing: Science Press, 2017: 256-257.

[18] ALQAHTANI N, KHAN W A, ALHUMAIDI M H, et al. Use of glycated hemoglobin in the diagnosis of diabetes mellitus and pre-diabetes and role of fasting plasma glucose, oral glucose tolerance test [J]. Int J Prev Med, 2013, 4(9): 1025-1029.

[19] 郭燕. 糖化血红蛋白与糖耐量检测在糖尿病诊治中的应用价值[J]. 医学信息, 2015(30): 317-318.
GUO Y. Application value of glycosylated hemoglobin and glucose tolerance testing in the diagnosis and treatment of diabetes mellitus[J]. Medical Information, 2015(30): 317-318.

[20] 张贤梅, 江波, 孙勤国. 糖尿病实验动物模型的研究进展[J]. 中西医结合研究, 2017, 9(2): 101-104.
ZHANG X M, JIANG B, SUN Q G. Research progress of experimental animal model of diabetes mellitus [J]. Journal of Chinese Integrative Medicine, 2017, 9(2): 101-104.

[21] 万玉玲, 张艳春, 彭白露,等. 中老年食蟹猴群体自发型糖尿病的筛选[J]. 动物学研究, 2011, 32(3): 307-310.
WAN Y L, ZHANG Y C, PENG B L, et al. Screening of spontaneous diabetes mellitus in middle-and old-aged cynomolgus monkey[J]. Zoological Research, 2011, 32(3): 307-310.

[22] 韦祝梅, 罗绍忠, 黄俊华, 等. 不同体重指数食蟹猴糖耐量试验初步探讨[J]. 医学信息, 2013(18): 162.
WEI Z M, LUO S Z, HUANG J H, et al. Preliminary study on the glucose tolerance test in different body mass index of Cynomolgus monkeys[J]. Medical Information, 2013(18): 162.

（编辑：薛泽玲）