(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109979599 A
(43)申请公布日 2019.07.05

(21)申请号 201711445663.4

(22)申请日 2017.12.27

(71)申请人 中国科学院沈阳自动化研究所
地址 110016 辽宁省沈阳市东陵区南塔街114号

(72)发明人 石刚 孙靖哲 赵伟 李虎阳 刘晓松

(74)专利代理机构 沈阳科苑专利商标代理有限公司 21002
代理人 李巨智

(51)Int.Cl.
*G16H 50/50*(2018.01)

权利要求书2页 说明书4页 附图1页

(54)发明名称

一种基于机器学习的糖尿病智能预测模型的建立方法

(57)摘要

本发明涉及一种基于机器学习的糖尿病智能预测模型的建立方法，从数据库调取个体生理参数数据，并将个体生理参数数据通过稀疏自编码器转化为xgboost模型可用的稀疏特征；将稀疏特征带入xgboost模型进行训练，在训练过程中调整xgboost模型的结构与参数，得到训练后的xgboost模型；将训练后的xgboost模型的参数进行提取，转化为预测模型，并根据预测模型得到预测结果。本发明采用稀疏自编码器自动提取有效特征，减少了对医学先验知识的需求；通过易采集的生理参数，在降低数据采集难度的情况下保证了预测准确率；可重复使用预测模型，不需要频繁再训练，降低了时间复杂度。


CN 109979599 A

1.一种基于机器学习的糖尿病智能预测模型的建立方法,其特征在于:包括以下步骤:

步骤1:从数据库调取个体生理参数数据,并将个体生理参数数据通过稀疏自编码器转化为xgboost模型可用的稀疏特征;

步骤2:将稀疏特征带入xgboost模型进行训练,在训练过程中调整xgboost模型的结构与参数,得到训练后的xgboost模型;

步骤3:将训练后的xgboost模型的参数进行提取,转化为预测模型,并根据预测模型得到预测结果。

2.根据权利要求1所述的基于机器学习的糖尿病智能预测模型的建立方法,其特征在于:所述个体生理参数数据包括静态信息和动态信息,其中

静态信息包括:年龄、性别、BMI、腰围、高血压家族史、糖尿病家族史、人口学测量指标、种族;

动态信息包括:收缩压、FPG、腰围、锻炼习惯、饮食习惯、体重变化情况、BMI变化情况、锻炼习惯变化情况。

3.根据权利要求1所述的基于机器学习的糖尿病智能预测模型的建立方法,其特征在于:所述稀疏自编码器为:

$$J_{sparse}(W,b) = J(W,b) + \beta \sum_{j=1}^{S_2} KL(\rho \parallel \widehat{\rho_j})$$

其中,W为自编码器网络权重,b为网络偏置,β为稀疏惩罚系数,ρ为理想活跃度,$\widehat{\rho_j}$为实际活跃度,$\sum_{j=1}^{S_2} KL(\rho \parallel \widehat{\rho_j})$表示了具有交叉熵性质的惩罚因子限制稀疏度,其公式为:

$$\sum_{j=1}^{S_2} KL(\rho \parallel \widehat{\rho_j}) = \sum_{j=1}^{S_2} \rho log \frac{\rho}{\widehat{\rho_j}} + (1-\rho) log \frac{1-\rho}{1-\widehat{\rho_j}}$$

其中,平均活跃度$\widehat{\rho_j}$的数学定义为:

$$\widehat{p_j} = \frac{1}{m} \sum_{i=1}^{m} [a_j^{(2)}(x^{(i)})]$$

这里m为神经元个数,$a_j^{(2)}$为隐藏层权重,$x^{(i)}$为神经元的值。

4.根据权利要求1所述的基于机器学习的糖尿病智能预测模型的建立方法,其特征在于:所述调整xgboost模型的结构包括总体结构优化和个体结构优化。

5.根据权利要求4所述的基于机器学习的糖尿病智能预测模型的建立方法,其特征在于:所述总体结构优化为:

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)) + \Omega(f_t) + constant$$

其中,$\Omega(f_t)$为L2正则项,l(.)为需要降低的残差,$y_i$为第i个基学习器输出的目标函数,n为基学习器个数,$f_t$为当前学习器的目标函数,constant为常数,评价标准为MSE。

6.根据权利要求4所述的基于机器学习的糖尿病智能预测模型的建立方法,其特征在

于:所述个体结构优化为通过贪心算法对叶子节点进行分割:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma$$

其中,$G_L$为左子树的增益,$G_R$为右子树的增益,$H_L$,$H_R$为左右子树的熵,λ为正则系数,$\frac{G_L^2}{H_L+\lambda}$为左子树的分数,$\frac{G_R^2}{H_R+\lambda}$为右子树的分数,$\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}$为不分裂的分数,γ为新叶子节点引入的复杂度代价。

7.根据权利要求1所述的基于机器学习的糖尿病智能预测模型的建立方法,其特征在于:需要调整的xgboost模型的参数包括行采样、列采样、树个数、树深度、学习率、正则化系数、叶子节点最大值。

8.根据权利要求1所述的基于机器学习的糖尿病智能预测模型的建立方法,其特征在于:所述预测模型为:

$$f(x) = f_M(x) + \sum_{j=1}^{J} \gamma_{jm} I$$

其中,M为基学习器数量,$\gamma_{jm}$为第m个基学习器的复杂度,J为特征总数,I为单位矩阵。

# 一种基于机器学习的糖尿病智能预测模型的建立方法

## 技术领域

[0001] 本发明涉及医学模型领域，具体地说是一种基于机器学习的糖尿病智能预测模型的建立方法。

## 背景技术

[0002] 早期在不采血的情况下进行诊断是糖尿病疾病临床诊断中的难题之一，如今，医学临床对非采血采尿的糖尿病临床诊断方法仍然十分有限，目前主要有一些形态学诊断方法，例如FINDRISK、ADRS和CDRS以及AUSDRISK等，不仅其采集方法复杂，而且诊断精度较低，绝大多数只能在病情发展到有明显病理改变时才能检测到。

## 发明内容

[0003] 针对现有技术的不足，本发明提供一种基于机器学习的糖尿病智能预测模型的建立方法，利用人工智能技术采集多个生理参数，并提取出生理参数中的有益信息形成特征，并以此训练模型，预测糖尿病。

[0004] 本发明为实现上述目的所采用的技术方案是：

[0005] 一种基于机器学习的糖尿病智能预测模型的建立方法，包括以下步骤：

[0006] 步骤1：从数据库调取个体生理参数数据，并将个体生理参数数据通过稀疏自编码器转化为xgboost模型可用的稀疏特征；

[0007] 步骤2：将稀疏特征带入xgboost模型进行训练，在训练过程中调整xgboost模型的结构与参数，得到训练后的xgboost模型；

[0008] 步骤3：将训练后的xgboost模型的参数进行提取，转化为预测模型，并根据预测模型得到预测结果。

[0009] 所述个体生理参数数据包括静态信息和动态信息，其中

[0010] 静态信息包括：年龄、性别、BMI、腰围、高血压家族史、糖尿病家族史、人口学测量指标、种族；

[0011] 动态信息包括：收缩压、FPG、腰围、锻炼习惯、饮食习惯、体重变化情况、BMI变化情况、锻炼习惯变化情况。

[0012] 所述稀疏自编码器为：

[0013] $$J_{sparse}(W,b) = J(W,b) + \beta \sum_{j=1}^{S_2} KL(\rho \parallel \hat{\rho}_j)$$

[0014] 其中，W为自编码器网络权重，b为网络偏置，β为稀疏惩罚系数，ρ为理想活跃度，$\hat{\rho}_j$为实际活跃度，$\sum_{j=1}^{S_2} KL(\rho \parallel \hat{\rho}_j)$表示了具有交叉熵性质的惩罚因子限制稀疏度，其公式为：

[0015] $$\sum_{j=1}^{S_2} KL(\rho \parallel \widehat{\rho_j}) = \sum_{j=1}^{S_2} \rho log \frac{\rho}{\widehat{\rho_j}} + (1-\rho) log \frac{1-\rho}{1-\widehat{\rho_j}}$$

[0016] 其中，平均活跃度$\widehat{\rho_j}$的数学定义为：

[0017] $$\widehat{p_j} = \frac{1}{m}\sum_{i=1}^{m}[a_j^{(2)}(x^{(i)})]$$

[0018] 这里m为神经元个数，$a_j^{(2)}$为隐藏层权重，$x^{(i)}$为神经元的值。

[0019] 所述调整xgboost模型的结构包括总体结构优化和个体结构优化。

[0020] 所述总体结构优化为：

[0021] $$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)) + \Omega(f_t) + constant$$

[0022] 其中，$\Omega(f_t)$为L2正则项，l(.)为需要降低的残差，$y_i$为第i个基学习器输出的目标函数，n为基学习器个数，$f_t$为当前学习器的目标函数，constant为常数，评价标准为MSE。

[0023] 所述个体结构优化为通过贪心算法对叶子节点进行分割：

[0024] $$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma$$

[0025] 其中，$G_L$为左子树的增益，$G_R$为右子树的增益，$H_L$，$H_R$为左右子树的熵，λ为正则系数，$\frac{G_L^2}{H_L+\lambda}$为左子树的分数，$\frac{G_R^2}{H_R+\lambda}$为右子树的分数，$\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}$为不分裂的分数，γ为新叶子节点引入的复杂度代价。

[0026] 需要调整的xgboost模型的参数包括行采样、列采样、树个数、树深度、学习率、正则化系数、叶子节点最大值。

[0027] 所述预测模型为：

[0028] $$f(x) = f_M(x) + \sum_{j=1}^{J} \gamma_{jm} I$$

[0029] 其中，M为基学习器数量，$\gamma_{jm}$为第m个基学习器的复杂度，J为特征总数，I为单位矩阵。

[0030] 本发明具有以下有益效果及优点：

[0031] 1.本发明采用稀疏自编码器自动提取有效特征，减少了对医学先验知识的需求；

[0032] 2.本发明通过易采集的生理参数，在降低数据采集难度的情况下保证了预测准确率；

[0033] 3.本发明通过可重复使用的预测模型，不需要频繁再训练，降低了时间复杂度。

## 附图说明

[0034] 图1是本发明的方法流程图。

## 具体实施方式

[0035] 下面结合附图及实施例对本发明做进一步的详细说明。

[0036] 如图1所示为本发明的方法流程图。

[0037] 一种基于机器学习的糖尿病智能预测模型，建立方法包括以下步骤：

[0038] 步骤1：从数据库调取个体生理参数数据，所述生理参数数据包括：年龄、性别、BMI、腰围、高血压家族史、糖尿病家族史、人口学测量指标、种族、收缩压、FPG、腰围、锻炼习惯、饮食习惯、体重变化情况、BMI变化情况、锻炼习惯变化情况。

[0039] 其中年龄、性别、BMI、腰围、高血压家族史、糖尿病家族史、人口学测量指标、种族为静态信息，收缩压、FPG、腰围、锻炼习惯、饮食习惯、体重变化情况、BMI变化情况、锻炼习惯变化情况为动态信息。

[0040] 步骤2：将生理数据通过自编码器转化为模型可用特征。

[0041] 步骤3：将特征带入模型进行训练，训练过程中调整模型的结构与参数。

[0042] 步骤4：将训练好的模型的参数进行提取，并转化为预测模型。

[0043] 步骤5：根据预测模型得到预测结果。预测模型的公式为：

[0044] $$f(x) = f_M(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$$

[0045] 其中，M为基学习器数量，γ为模型复杂度。

[0046] 采用上述方法，通过分析数据库中采集的个体生理参数信息，结合基于机器学习的糖尿病智能预测模型计算就能得到新个体是否为糖尿病患者的信息，将该信息存储到数据库中，该数据库就可以用作人群健康程度的大数据检测。

[0047] 更进一步的，步骤2采用稀疏自编码器对生理参数数据进行稀疏特征化：

[0048] 步骤1：保证模型特征的稀疏性，加入平均活跃度的定义：

[0049] $$\hat{p}_j = \frac{1}{m}\sum_{i=1}^{m}[a_j^{(2)}(x^{(i)})]$$

[0050] 步骤2：选择具有交叉熵性质的惩罚因子限制稀疏度：

[0051] $$\sum_{j=1}^{S_2} \rho log \frac{\rho}{\hat{\rho}_j} + (1-\rho) log \frac{1-\rho}{1-\hat{\rho}_j}$$

[0052] 其中，$S_2$是隐藏层中隐藏神经元的数量，而索引j依次代表隐藏层中的每一个神经元。

[0053] 步骤3：优化带有惩罚因子的损失函数：

[0054] $$J_{sparse}(W,b) = J(W,b) + \beta \sum_{j=1}^{S_2} KL(\rho \| \hat{\rho}_j)$$

[0055] 采用上述步骤，稀疏自编码器是运用很广泛的特征提取技术，能够满足个体生理信号从稠密到稀疏特征的转换，并使得特征维度更高，便于模型对于特征的利用和处理。

[0056] 更进一步的，步骤3将特征带入模型进行训练，训练过程中调整模型的结构：

[0057] 步骤①：这里的模型选择使用xgboost，其需要优化的目标函数为：

[0058] $$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)) + \Omega(f_t) + constant$$

[0059] 其中Ω($f_t$)为L2正则项，l(.)为需要降低的残差。

[0060] 步骤②：采用MSE作为评价标准优化目标函数，公式为：

[0061] $\text{minimize } \frac{1}{2}(y-y')^2$

[0062] 这里y为真实值，y′为模型的预测值。

[0063] 采用上述步骤，能从模型的结构角度降低预测的经验误差，发现数据的签字信息，提升预测准确率。

[0064] 更进一步的，步骤①的模型参数主要包括行采样、列采样、树个数、树深度、学习率、正则化系数、叶子节点最大值。

[0065] 采用上述方法，能够从模型的参数调整角度来提升模型整体的泛化能力，是模型产生的函数更接近于数据在假设空间的真实分布。

[0066] 更进一步的，步骤4将训练好的模型的参数进行提取，并转化为预测模型：采用pickle将训练好的智能预测模型的参数进行参数服务器保存，以便在进行新数据预测时调用。

[0067] 采用上述方法，能够将训练好的模型参数保存在分布式服务器中，一方面避免了模型做预测需要再训练的麻烦，另一方面分布式服务器可以降低因为硬件损坏造成数据损失的可能性。

[0068] 更进一步的，步骤5中调用参数服务器中存储的参数，形成预测模型，对新个体的患病与否给出准确的预测，并实时动态调整模型参数，进行参数更新和覆盖，预测公式为：

[0069] $f(x) = f_M(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$

[0070] 其中，M为基学习器数量，γ为模型复杂度。

[0071] 采用上述方法，能够在预测中实时调用训练好的参数，加速预测速度，提升预测准确率，并可以根据模型随时间衰减的函数，适当对模型进行重新训练，以及覆盖原参数。

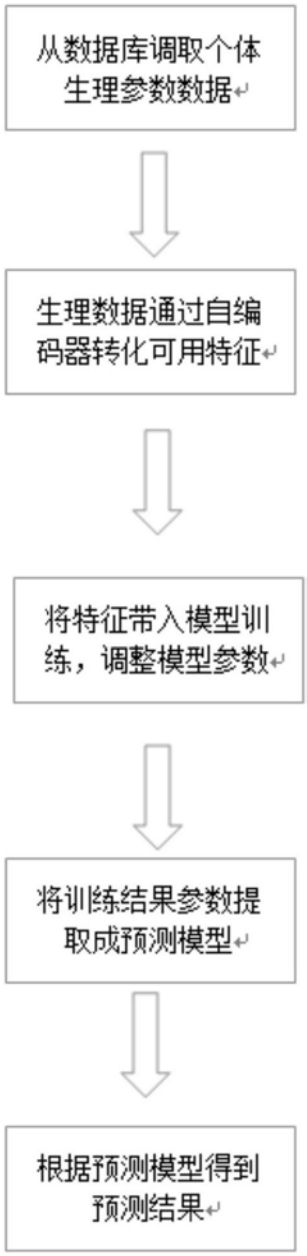
从数据库调取个体生理参数数据
生理数据通过自编码器转化可用特征
将特征带入模型训练，调整模型参数
将训练结果参数提取成预测模型
根据预测模型得到预测结果

图1