(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 106682412 A
(43)申请公布日 2017.05.17

(21)申请号 201611199219.4

(22)申请日 2016.12.22

(71)申请人 浙江大学
地址 310013 浙江省杭州市西湖区余杭塘路866号

(72)发明人 吴健 周立水 顾盼 邱奇波 邓水光 李莹 尹建伟 吴朝晖

(74)专利代理机构 杭州天勤知识产权代理有限公司 33224
代理人 胡红娟

(51)Int.Cl.
G06F 19/00(2011.01)

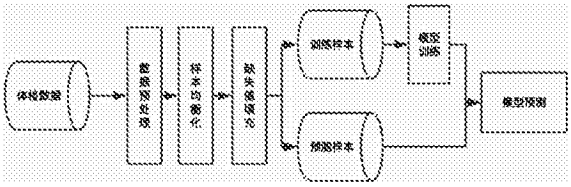
权利要求书1页 说明书4页 附图3页

(54)发明名称

一种基于医疗体检数据的糖尿病预测方法

(57)摘要

本发明公开了一种基于医疗体检数据的糖尿病预测方法，包括：(1)对每个用户的体检数据进行处理，得到完整的体检数据；(2)将完整的患糖尿病的体检数据作为正训练样本，将完整的未患糖尿病的体检数据作为负训练样本；采用GBDT+LR模型进行训练，并根据模型的效果进行模型调整融合，得到最终预测模型；(3)将处理后的新用户的体检数据作为预测样本输入到最终预测模型，得到新用户的患糖尿病概率。通过该方法可以辅助医生进行更好的判断，病人更好的了解自身患病的风险。



CN 106682412 A

1.一种基于医疗体检数据的糖尿病预测方法,包括以下步骤:

(1)对每个用户的体检数据进行处理,得到完整的体检数据;

(2)将完整的患糖尿病的体检数据作为正训练样本,将完整的未患糖尿病的体检数据作为负训练样本;采用GBDT+LR模型进行训练,并根据模型的效果进行模型调整融合,得到最终预测模型;

(3)将处理后的新用户的体检数据作为预测样本输入到最终预测模型,得到新用户的患糖尿病概率。

2.根据权利要求1所述基于医疗体检数据的糖尿病预测方法,其特征在于:步骤1的具体步骤为:

(1-1)对每个用户的体检数据进行预处理,得到同一格式的体检数据;

(1-2)对同一格式的体检数据进行均衡化,得到均衡化的体检数据;

(1-3)对均衡化的体检数据进行数据缺失值填充,得到完整的体检数据。

3.根据权利要求2所述基于医疗体检数据的糖尿病预测方法,其特征在于:在步骤(1-1)中,进行体检数据预处理的过程为:首先,对体检数据中原生的诊断结果、体检项目名称以及体检项目结果,采用自然语言处理方法进行分析,得到分析结果;然后,对分析结果进一步地清洗和标准化,转换为同一格式的体检数据。

4.根据权利要求2所述基于医疗体检数据的糖尿病预测方法,其特征在于:对体检数据进行均衡化的方法有三种,分别为:

(a)重采样法:通过重复采样患糖尿病的体检数据以扩大患糖尿病的体检数据的数量;以达到正负例训练样本的均衡;

(b)欠采样法:通过少量采样未患糖尿病的体检数据以缩小未患糖尿病的体检数据的数量,以达到正负例训练样本的均衡;

(c)权值调整法:通过改变患糖尿病的体检数据与未患糖尿病的体检数据的权值比例以使得正负例训练样本的总权值一致,以达到正负例训练样本的均衡。

5.根据权利要求4所述基于医疗体检数据的糖尿病预测方法,其特征在于:对体检数据进行均衡化的方法为:采用重采样法与欠采样法结合的方式,随机采样正例患糖尿病的体检数据,并排序抽取缺失数据较少的负例未患糖尿病的体检数据。

6.根据权利要求2所述基于医疗体检数据的糖尿病预测方法,其特征在于:在步骤(1-3)中,采用kNN算法对体检数据缺失值进行预测,利用用户自身其他特征寻找最相似的k个用户,综合k个用户的相似性加权平均值进行体检数据的缺失值填充,k为用户的个数。

一种基于医疗体检数据的糖尿病预测方法

技术领域

[0001] 本发明属于大数据医疗领域，具体涉及一种医疗体检数据的糖尿病预测方法。

背景技术

[0002] 随着人们生活水平的提高、保健意识的增强，健康体检逐渐成为一种社会时尚，人们已经改变了只有在得病时才去医院的传统观念，定期体检已经被大多数人所接受。因此，医院积累了海量的电子体检数据，使大数据有了用武之地。

[0003] 大数据医疗是当前的一个热点，是指通过大数据相关技术，分析医疗领域的数据并挖掘其中的知识从而大幅度提高医疗服务。在过去的几十年中，大数据已经深深地影响了每一个企业，包括医疗保健行业。如今，大量的数据可以让医疗保健更加高效，更加个性化。

[0004] 今年，世界卫生组织(WHO)发出警告，我国约有1.1亿名糖尿病患者，约占中国成年人总数的1/10。若不尽快采取行动，减少不健康饮食和缺乏运动等生活方式中的危险因素，预计该数字将在2040年增至1.5亿人，给民众健康和社会经济带来严重影响。糖尿病除了对患者及其家人朋友造成身心的伤害，也带来巨大的经济损失。我国每年投入近1734亿人民币(250亿美元)用于糖尿病管理；用于糖尿病的直接医疗支出占中国医疗支出的13％。这些数据还未包括糖尿病相关疾病给患者家庭和公司带来的经济损失。将大数据引入糖尿病医疗领域，不但能减小医生压力，还能让病人平时过得更舒服。

发明内容

[0005] 鉴于上述，本发明提供了一种基于医疗体检数据的糖尿病预测方法，该方法是通过分析体检数据中病人的各项数据指标和医生对病人体检数据的诊断，建立体检数据和体检诊断之间的关联，预测病人是否可能患糖尿病，从而辅助医生进行更好的判断，令病人更好的了解自身患病的风险。

[0006] 一种基于医疗体检数据的糖尿病预测方法，包括以下步骤：

[0007] (1)对每个用户的体检数据进行处理，得到完整的体检数据；

[0008] (2)将完整的患糖尿病的体检数据作为正训练样本，将完整的未患糖尿病的体检数据作为负训练样本；采用GBDT+LR模型进行训练，并根据模型的效果进行模型调整融合，得到最终预测模型；

[0009] (3)将处理后的新用户的体检数据作为预测样本输入到最终预测模型，得到新用户的患糖尿病概率。

[0010] 步骤1的具体步骤为：

[0011] (1-1)对每个用户的体检数据进行预处理，得到同一格式的体检数据；

[0012] (1-2)对同一格式的体检数据进行均衡化，得到均衡化的体检数据；

[0013] (1-3)对均衡化的体检数据进行数据缺失值填充，得到完整的体检数据。

[0014] 在步骤(1-1)中，进行体检数据预处理的过程为：首先，对体检数据中原生的诊断

结果、体检项目名称以及体检项目结果，采用自然语言处理方法进行分析，得到分析结果；然后，对分析结果进一步地清洗和标准化，转换为同一格式的体检数据，使更多资料可用。

[0015] 在步骤(1-2)中，由于在体检数据中，患糖尿病的用户只占据其中的一部分，因此，通过扩大患糖尿病的体检数据(小样本)，缩小未患糖尿病的体检数据(大样本)的方法，得到数量相等的正负例训练样本，以达到正负例样本的均衡化，便于后续的模型使用。

[0016] 数据样本均衡化：在分类问题中，经常会遇到正负例样本数据量不等的情况，比如正例样本为10w条数据，负例样本只有1w条数据，此时需要进行样本的均衡化，使得正负例样本达到平衡。

[0017] 对体检数据进行均衡化的方法有三种，分别为：

[0018] (a)重采样法：通过重复采样患糖尿病的体检数据以扩大患糖尿病的体检数据的数量；以达到正负例训练样本的均衡。

[0019] (b)欠采样法：通过少量采样未患糖尿病的体检数据以缩小未患糖尿病的体检数据的数量，以达到正负例训练样本的均衡。

[0020] (c)权值调整法：通过改变患糖尿病的体检数据与未患糖尿病的体检数据的权值比例以使得正负例训练样本的总权值一致，以达到正负例训练样本的均衡。

[0021] 作为优选，采用重采样法与欠采样法结合的方式，即随机采样正例患糖尿病的体检数据，并排序抽取缺失数据较少的负例未患糖尿病的体检数据。这样既扩大了正例样本数据量，又筛选了较差的负例样本。

[0022] 在步骤(1-3)中，数据值缺失是指在数据获取过程中因为自然原因和人为原因导致数据不完整，体检数据中同样也存在数据值缺失的情形，因此，需要对体检数据进行缺失值填充。进行缺失值填充的方法包括三种，分别为：

[0023] (a)直接删除法：直接删除有缺失数据的体检数据。

[0024] (b)计算样本数据填充法：通过计算体检数据的中位数、众数、平均数以及随机分布值等，填充体检数据中的缺失值。

[0025] (c)综合整个样本数据填充法：找到最相似的体检数据，利用其进行体检数据的缺失值填充，或将缺失特征值映射高维空间。

[0026] 作为优选，本发明采用综合整个样本数据填充法进行数据缺失值填充，具体为：采用K最近邻(k-Nearest Neighbor,kNN)算法对体检数据缺失值进行预测，利用用户自身其他特征寻找最相似的k个用户，综合k个用户的相似性加权平均值进行体检数据的缺失值填充，k为用户的个数。

[0027] 在步骤(2)中，采用GBDT(Gradient Boosting Decision Tree)与LR(Logistic Regression)模型进行训练，并根据模型的效果进行模型调整融合，得到最终的模型。

[0028] 在步骤(3)中，首先，采用步骤(1-1)～步骤(1-3)对每个新用户的体检数据进行处理，然后，将处理后的新用户的体检数据作为预测样本输入到最终预测模型，得到新用户的患糖尿病概率。

[0029] 本发明基于医疗体检数据的糖尿病预测方法是通过分析用户的体检数据，利用大数据分析的手段，判断用户的患糖尿病病风险。从而促进各类糖尿病医疗应用的发展，不仅为医生的快速判断提供辅助依据，同时使病人对自身的潜在隐患有更直观的了解，具有的优点如下：

[0030] (1)对医疗体检数据进行预处理，将更多可用的体检数据转化为标准数据，不仅增多了训练样本，也能为更加复杂的体检数据提供预测服务。

[0031] (2)结合体检数据的特殊性，在扩大最终样本数量的同时，对低质量的样本进行了筛选。

[0032] (3)使用了KNN算法填充缺失数值，并局部调整优化，既可以利用已有数据进行推测，又不耗费过多计算资源。

[0033] (4)采用了GBDT+LR模型，既节省了人工处理分析特征的环节，又增强了非线性预测能力。

附图说明

[0034] 图1为本发明基于医疗体检数据的糖尿病预测方法的结构图；

[0035] 图2为医疗体检数据的诊断结果清洗与标准化示意图；

[0036] 图3为医疗体检数据的体检项目名称和结果清洗与标准化示意图；

[0037] 图4为正负例糖尿病体检者数据均衡化示意图；

[0038] 图5为数据缺失值填充方法分析图；

[0039] 图6为部分体检数据糖尿病预测结果图。

具体实施方式

[0040] 为了更为具体地描述本发明，下面结合附图及具体实施方式对本发明的技术方案进行详细说明。

[0041] 如图1所示，本发明基于医疗体检数据的糖尿病预测方法的具体步骤如下：

[0042] 步骤1，数据预处理：对每个用户的体检数据进行预处理，得到同一格式的体检数据。

[0043] 如图2所示，对于原生的医生诊断数据，由于出自不同医生和具体的不同场景，产生的诊断结果是复杂的，不能直接使用。例如所需要判断的糖尿病诊断中有糖尿病、糖尿病性视网膜病变、高度糖尿病发病风险等，需要经过数据清洗和标准化才能够作为诊断的标签使用。经过自然语言处理后，首先获取所有和糖尿病有关的诊断词，经过人工评判和相关医学知识的辅助，最终分成三个标签：糖尿病，疑似糖尿病，非糖尿病。同时，还有相关体检项目名称也需要进行清洗和标准化，如图3所示。比如说糖化血红蛋白项目，可能有糖化血红蛋白A1、糖化血红蛋白A1(HbA1)、糖化血红蛋白、A1(HbA1)等，它们都是指代同一个体检项目，只是在不同体检套餐中有不同的名称。除了体检项目名称，还有体检项目结果也需要进行清洗标准化，例如：结果可能为拒检、拒测、未检、21、88cm、左手：135右手：129、76未、++32等，这些数据都会在清洗和标准化后有统一的数据格式和单位。

[0044] 步骤2，数据样本均衡化：对同一格式的体检数据进行均衡化，得到均衡化的体检数据。

[0045] 如图4，在数据标准化之后，可以获取到一定数量的样本数据，此事往往存在的问题是，正负例样本不均衡，因为患糖尿病的用户在所有的体检用户中只是占据了一个部分。为均衡化正负例样本数据，本实施例采用的是随机采样正例患糖尿病用户的数据，并排序抽取缺失数据较少的负例样本数据。这样既扩大了正例样本数据量，又筛选了较差的负例

样本。

[0046] 步骤3，缺失值填充：对均衡化的体检数据进行数据缺失值填充，得到完整的体检数据。

[0047] 在均衡化的数据中，仍然存在许多数据缺失值，需要进行填充，如图5。本实施例中选择采用其他特征对缺失值进行预测。对于简单计算数据的中值、均值等进行填充的方法，存在随机性较大，会人为增加噪音的问题，会降低数据的准确性。而对于把缺失特征值映射到高维空间的方法，则会增加计算量，需要较大的资源。选择的采用其他特征对缺失值进行预测的方法，主要需要依赖其他变量的相关性，对于体检数据而言较为适合。具体采用KNN算法计算该数据最相似的k条记录，根据其相似性加权平均获得最终的填充值。也就是利用用户自身其他特征寻找最相似的k个用户，并综合k个用户的值进行体检数据的缺失值填充。

[0048] 步骤4，模型训练：将完整的患糖尿病的体检数据作为正训练样本，将完整的未患糖尿病的体检数据作为负训练样本；采用GBDT+LR模型进行训练，并根据模型的效果进行模型调整融合，得到最终预测模型。

[0049] GBDT又叫MART(Multiple Additive Regression Tree)，是一种常用的非线性模型，它基于集成学习中的boosting思想，每次迭代都在减少残差的梯度方向新建立一颗决策树，迭代多少次就会生成多少颗决策树。GBDT的思想使其具有天然优势可以发现多种有区分性的特征以及特征组合，决策树的路径可以直接作为LR输入特征使用，省去了人工寻找特征、特征组合的步骤。LR是一种线性拟合模型，可以利用Logistic函数(或称为Sigmoid函数)变成分类器。

[0050] 步骤5，模型预测：将处理后的新用户的体检数据作为预测样本输入到最终预测模型，得到新用户的患糖尿病概率。

[0051] 获取到模型结果后，对于每一个新的用户体检数据，只需要自动化上述流程即可得到他的患糖尿病概率。如图6所示的是部分体检数据糖尿病预测结果图，分析从图6可得：利用该方法进行预测得到的糖尿病预测准确率很好。

[0052] 以上所述的具体实施方式对本发明的技术方案和有益效果进行了详细说明，应理解的是以上所述仅为本发明的最优选实施例，并不用于限制本发明，凡在本发明的原则范围内所做的任何修改、补充和等同替换等，均应包含在本发明的保护范围之内。

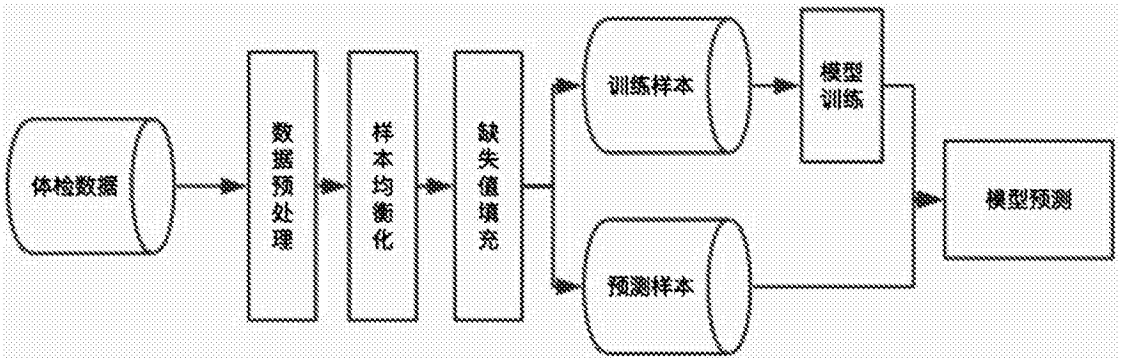

体检数据
数据预处理
样本均衡化
缺失值填充
训练样本
预测样本
模型训练
模型预测

图1

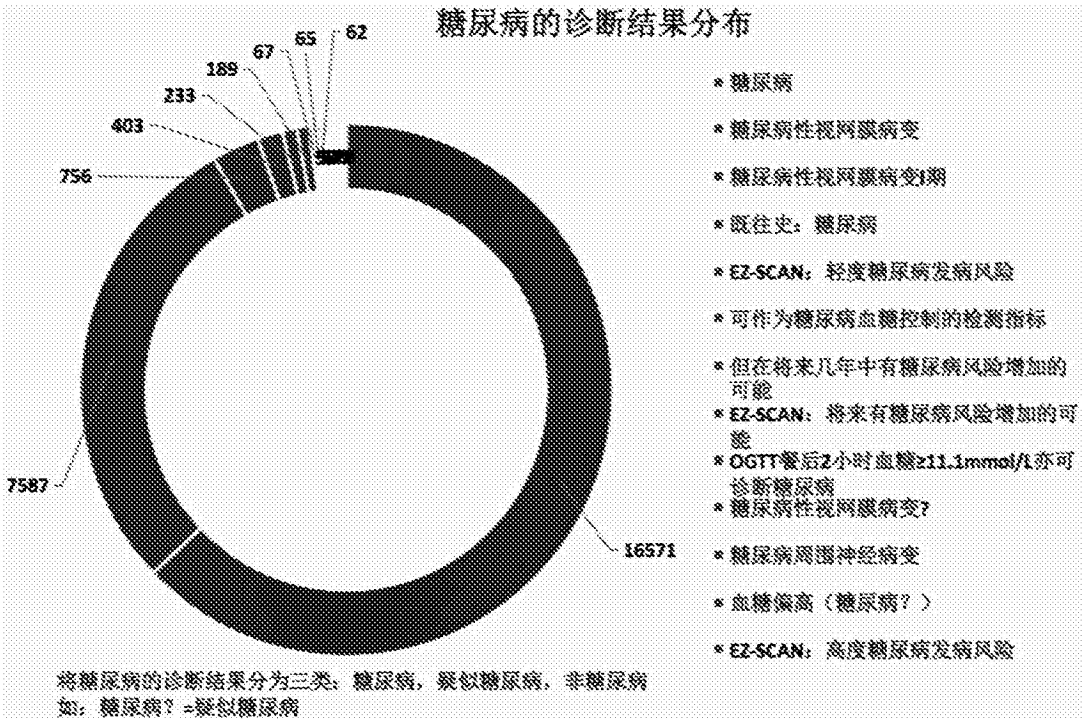

糖尿病的诊断结果分布
62
65
67
189
233
403
756
7587
16571
* 糖尿病
* 糖尿病性视网膜病变
* 糖尿病性视网膜病变1期
* 既往史：糖尿病
* EZ-SCAN：轻度糖尿病发病风险
* 可作为糖尿病血糖控制的检测指标
* 但在将来几年中有糖尿病风险增加的可能
* EZ-SCAN：将来有糖尿病风险增加的可能
* OGTT餐后2小时血糖≥11.1mmol/L亦可诊断糖尿病
* 糖尿病性视网膜病变?
* 糖尿病周围神经病变
* 血糖偏高（糖尿病？）
* EZ-SCAN：高度糖尿病发病风险
将糖尿病的诊断结果分为三类：糖尿病，疑似糖尿病，非糖尿病
如：糖尿病？=疑似糖尿病

图2

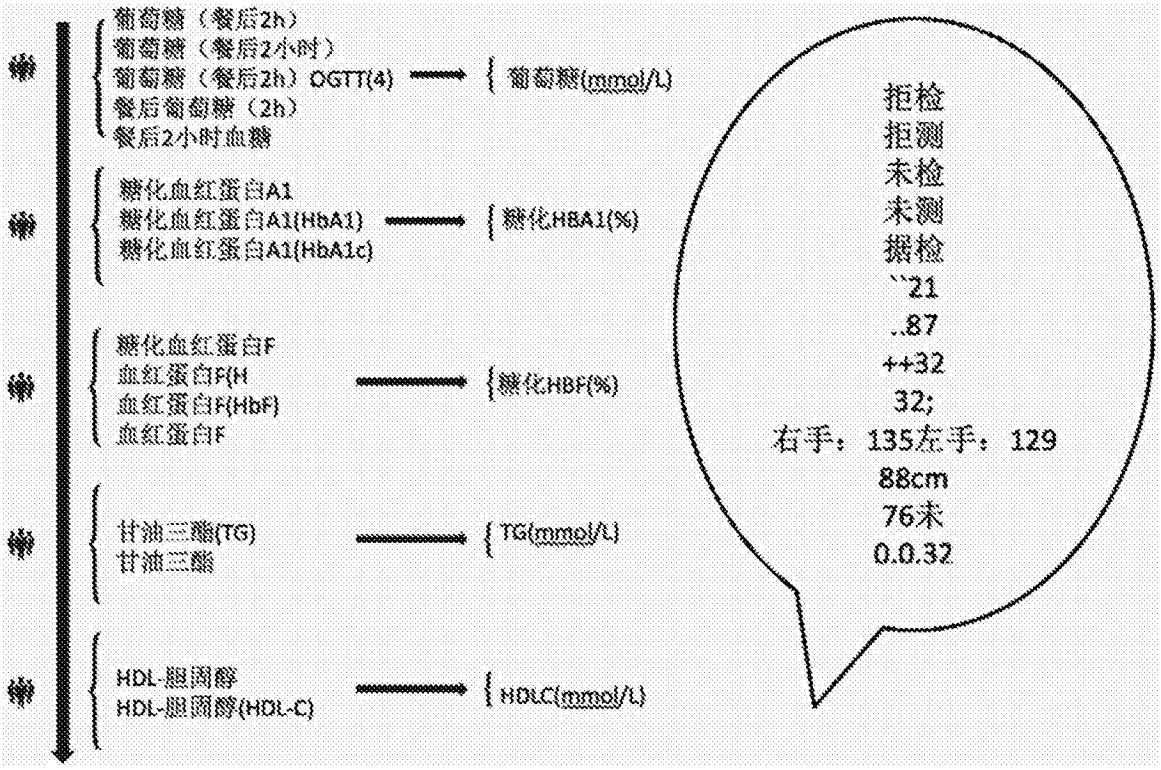

葡萄糖（餐后2h）
葡萄糖（餐后2小时）
葡萄糖（餐后2h）OGTT(4)
餐后葡萄糖（2h）
餐后2小时血糖
葡萄糖(mmol/L)
糖化血红蛋白A1
糖化血红蛋白A1(HbA1)
糖化血红蛋白A1(HbA1c)
糖化HBA1(%)
糖化血红蛋白F
血红蛋白F(H
血红蛋白F(HbF)
血红蛋白F
糖化HBF(%)
甘油三酯(TG)
甘油三酯
TG(mmol/L)
HDL-胆固醇
HDL-胆固醇(HDL-C)
HDLC(mmol/L)
拒检
拒测
未检
未测
据检
``21
.87
++32
32;
右手：135左手：129
88cm
76未
0.0.32

图3

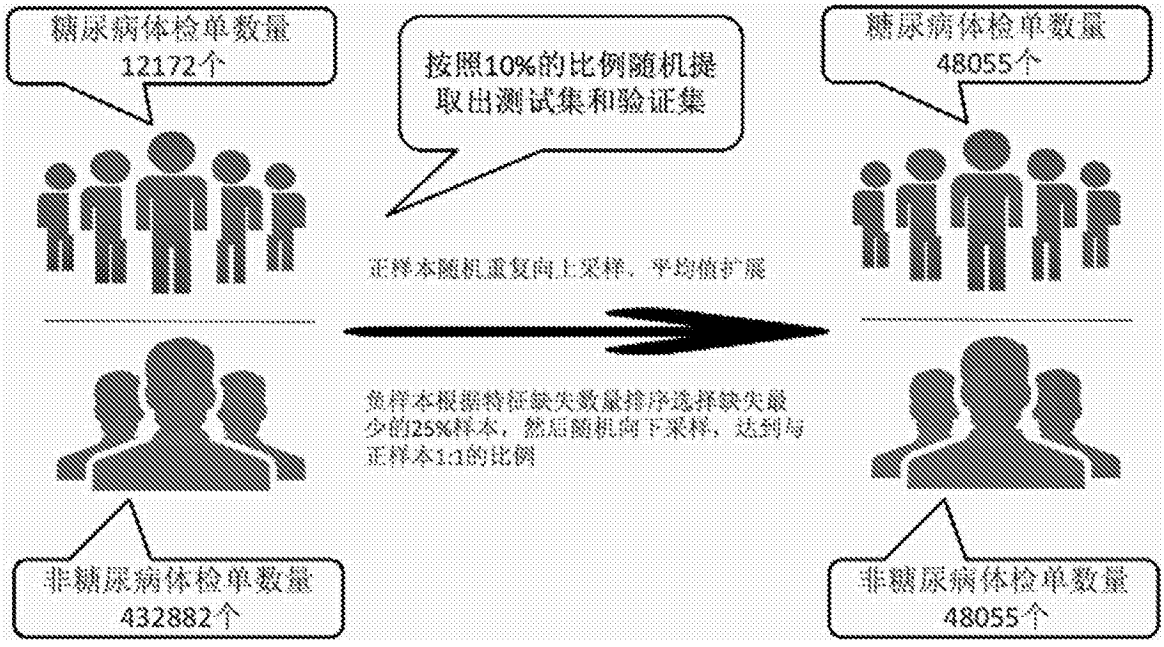

糖尿病体检单数量
12172个
按照10%的比例随机提取出测试集和验证集
糖尿病体检单数量
48055个
正样本随机重复向上采样，平均倍扩展
负样本根据特征缺失数量排序选择缺失最少的25%样本，然后随机向下采样，达到与正样本1:1的比例
非糖尿病体检单数量
432882个
非糖尿病体检单数量
48055个

图4

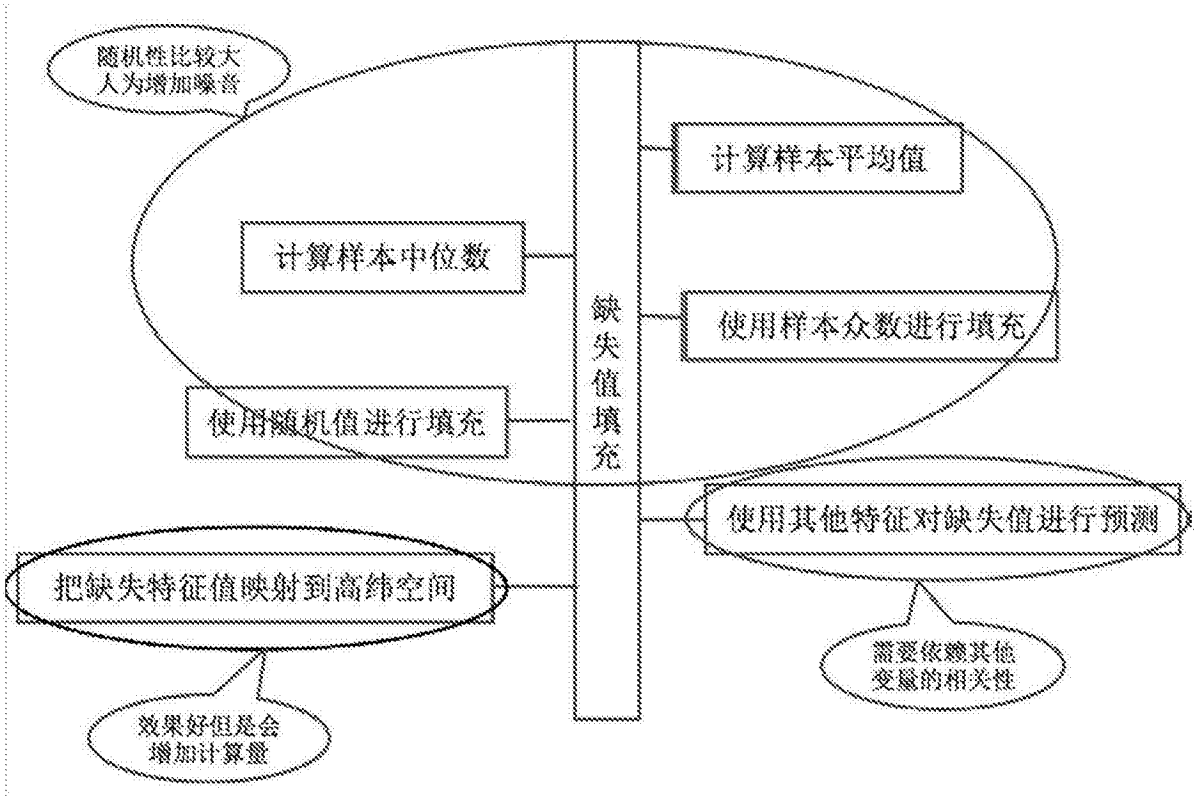

随机性比较大
人为增加噪音
计算样本平均值
计算样本中位数
缺失值填充
使用样本众数进行填充
使用随机值进行填充
使用其他特征对缺失值进行预测
把缺失特征值映射到高纬空间
需要依赖其他变量的相关性
效果好但是会增加计算量

图5

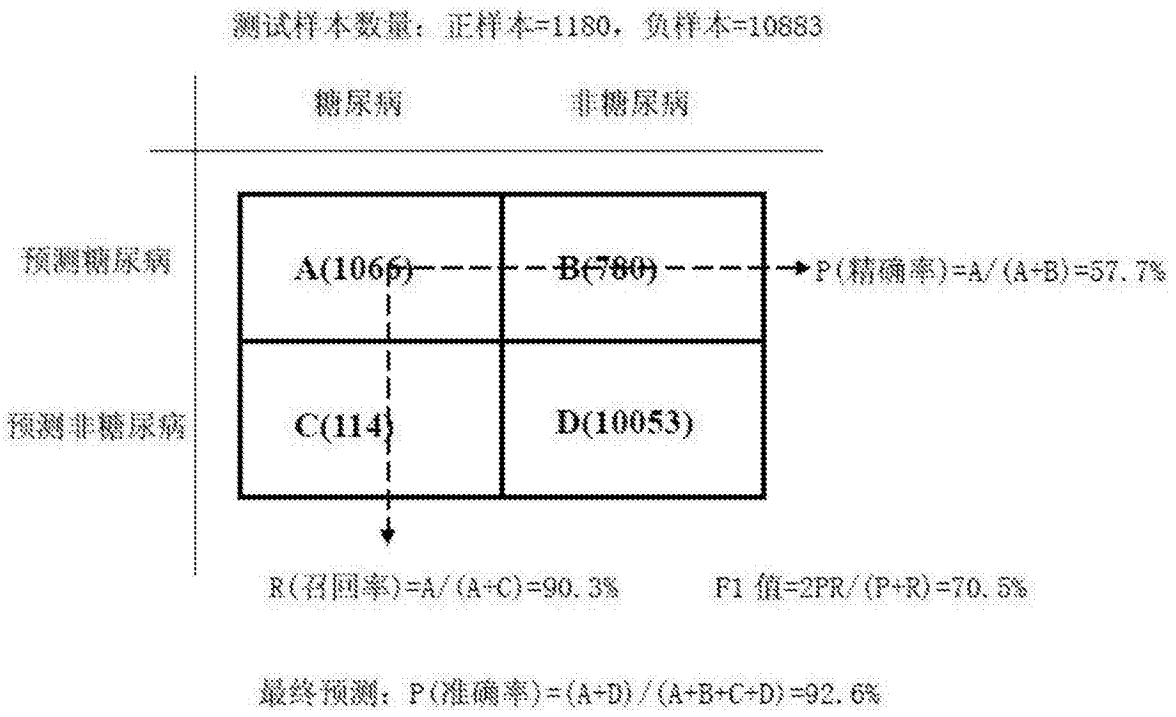

测试样本数量：正样本=1180，负样本=10883
糖尿病
非糖尿病
预测糖尿病
预测非糖尿病
A(1066)
B(780)
C(114)
D(10053)
P(精确率)=A/(A+B)=57.7%
R(召回率)=A/(A+C)=90.3%
F1 值=2PR/(P+R)=70.5%
最终预测：P(准确率)=(A+D)/(A+B+C+D)=92.6%

图6