(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 107403072 A
(43)申请公布日 2017.11.28

(21)申请号 201710665605.6

(22)申请日 2017.08.07

(71)申请人 北京工业大学
地址 100124 北京市朝阳区平乐园100号

(72)发明人 杨胜齐 吴寒 丁梦 王冰笛

(74)专利代理机构 北京思海天达知识产权代理有限公司 11203
代理人 沈波

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

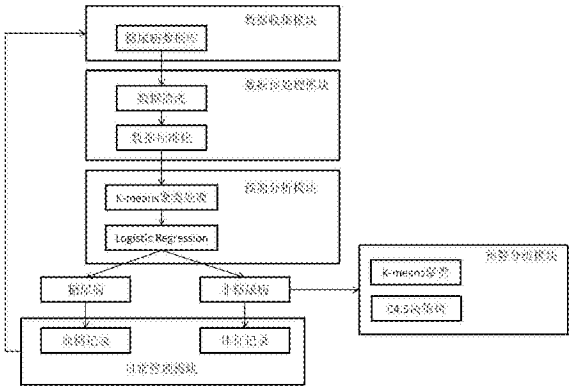
权利要求书3页 说明书5页 附图2页

(54)发明名称

一种基于机器学习的2型糖尿病预测预警方法

(57)摘要

本发明公开了一种基于机器学习的2型糖尿病预测预警方法，通过K-means算法和Logistic Regression算法建立先聚类再分类的糖尿病双层预测分析模型，对非糖尿病的分类结果通过C4.5算法和K-means算法分析出的规则进行预警分级，对糖尿病数据集通过日常数据收集管理后并进行更新维护以优化原始糖尿病双层预测分析模型。本方法包括数据收集模块、数据预处理模块、预测分析模块、预警分级模块和日常管理模块等五大模块。主要应用在糖尿病早期预测分析、高危人群预警分级以及糖尿病日常健康管理等三大方面。本方法在实际应用中具有更高的预测准确性，更加专注每个实例的健康参数，同时针对非糖尿病实例进行预警分级，能起到更好的防范作用，做到疾病的提早预防。



CN 107403072 A

1.一种基于机器学习的2型糖尿病预测预警方法，该方法通过K-means算法和Logistic Regression算法建立先聚类再分类的糖尿病双层预测分析模型，对非糖尿病的分类结果通过C4.5算法和K-means算法分析出的规则进行预警分级，对糖尿病数据集通过日常数据收集管理后并进行更新维护以优化原始糖尿病双层预测分析模型；本方法包括数据收集模块、数据预处理模块、预测分析模块、预警分级模块和日常管理模块，数据收集模块与数据预处理模块连接，数据预处理模块与预测分析模块连接，预测分析模块与预警分级模块连接，预警分级模块和日常管理模块连接；

其特征在于：本方法包括以下步骤：

(1)基于现有的健康大数据，获取医院、社康、体检中心等医疗单位内与糖尿病相关的健康数据以建立糖尿病参数数据库，健康数据包括年龄、身高、体重、腰围、臀围、收缩压、舒张压、心率、血糖、血氧、睡眠质量和饮食习惯等；首先对每一个实例以糖尿病和非糖尿病进行标记；

(2)对糖尿病参数数据库中的原始数据做数据预处理；预处理包括确定统一的数据项以及每一项的具体格式，具体格式为枚举型或数值型等，然后通过数据清洗和数据标准化提高原始数据的质量；

(3)使用K-means算法和Logistic Regression算法的双层预测分析模型分析处理过的数据集，对每个实例属于糖尿病或是非糖尿病进行预测分类；

(4)使用C4.5算法和K-means算法结合的分析模型结合现有高危人群划分标准对非糖尿病进行预警分级，提出无风险、低风险和高风险三类标识；

(5)通过引入最新的糖尿病参数数据，对所有糖尿病参数数据以统一规范的数据格式进行存储并及时反馈至数据收集模块以进行进一步优化；在丰富数据集的基础上，反复步骤(2)、(3)、(4)进行训练以优化本方法的实际应用效果；

所述的数据收集模块是基于现有的健康大数据，通过获取医院、社康、体检中心等医疗单位内与糖尿病相关的健康数据以建立糖尿病参数数据库，对每一个实例以糖尿病和非糖尿病进行标记；

所述的数据预处理模块分为两个基本步骤，分别是数据清洗和数据标准化；数据清洗包括一致性检查和缺失值处理，一致性检查是根据每个变量的合理取值范围和相互关系，检查数据是否合乎要求，发现超出正常范围、逻辑上不合理或者相互矛盾的数据；缺失值处理即将数据集中缺少的值使用平均值进行替代；数据标准化是在数据清洗完成后，为避免计算过程中数值复杂度，以及避免大数值区间的属性过分支配小数值区间的属性，将所有糖尿病参数中的属性进行Normalize，规范到数值区间[0,1]，使用公式：$\mathrm{Value}=\frac{value-x'}{s}$，其中x'表示糖尿病参数属性的平均值，s表示糖尿病参数属性的标准偏差，Value即糖尿病参数原属性值value进行标准化计算得到的结果；

所述的预测分析模块分为两部分，分别是K-means聚类处理和LogisticRegression模型处理；使用K-means聚类算法对剔除了分类标签结果的数据集进行一级处理，设定聚类数目为2，将结果与原始数据集进行对比，剔除聚类错误的数据项，以提供更加准确的数据集进入下一阶段的处理；使用LogisticRegression模型对上述处理过的数据集进行有监督的分类处理，分析结果可得预测准确性；

预警分级模块针对糖尿病数据集中的参数，根据C4.5算法和K-means算法基于现有数据分析糖尿病风险层级的规则并与糖尿病分级标准进行对比，以制定更加直接有效的预警分级标准；在此基础上，针对新数据的录入通过新的分级标准进行风险层级判断；

所述的日常管理模块包含血糖记录和体征记录；血糖记录针对糖尿病用户，通过引入最新的血糖记录针记录血糖参数数据以更新糖尿病数据库；体征记录针对所有用户，管理的糖尿病参数包括心率、血压、血氧、身高体重和腰臀比数据；所有参数均以统一规范的数据格式进行存储并及时反馈至数据收集模块进行优化。

2.根据权利要求1所述的一种基于机器学习的2型糖尿病预测预警方法，其特征在于：预测分析模块和预警分级模块是本方法的两大核心模块；

所述的预测分析模块由K-means算法和LogisticRegression算法组成；

K-means算法是典型的基于距离的聚类算法，采用距离作为相似度的度量指标，即规定对象间的距离值越小，其相似度越大；K-means算法所产生的簇都是由距离相近的对象组成，故其最终目标是找到这些紧凑且独立的簇；在K-means算法中，K值代表的是初始聚类中心的个数，聚类中心即簇，故K值的选取对聚类结果影响大；

针对包含n个糖尿病实例的初始数据集，n为糖尿病实例的个数且取正整数，对应的算法过程如下：

1)在给出数据集中的n个糖尿病实例里任意选取2个对象作为初始聚类中心；由于最终的分类结果为两类，将K值定为2；

2)对剩余的每个糖尿病实例分别计算与每个中心的距离，并根据剩余的每个糖尿病实例与各个簇中心的距离把剩余的每个糖尿病实例归到最近的中心的簇；

3)重新计算每个聚类的中心，判断中心是否发生变化；

4)循环步骤2)～3)步直至新的中心与原中心相等或小于指定阈值，即已收敛，则算法结束；通过误差函数判断收敛：$E=\sum_{i=1}^{k}\sum_{x\epsilon C_i} d(x,\mu(C_i))$，其中x表示糖尿病参数中的每个实例，$\mu(C_i)$表示聚类$C_i$的中心，$d(x,\mu(C_i))$表示x和$\mu(C_i)$之间的欧几里德距离，k的值为2；

在对糖尿病数据集使用K-means算法进行分析时，选取K值为2，将分析结果与数据集原始的分类标签进行比较，剔除错误的噪声数据，将剩余的数据集作为下一级LogisticRegression算法的输入；由于起初K-means算法的Seed值是随机选取的，在剔除数据过程中可能造成错误聚类数量过大，故在每次聚类分析结束后计算数据集剩余比；若比值大于75%，则进入下一步；若比值小于75%，则进入循环重新选取新的Seed值开始聚类；

以此方法降低人工选取Seed值导致错误分析的风险，并且能够有效控制原始数据集的不必要损失；

对高质量的数据集使用LogisticRegression算法进入第二级处理；

Logistic回归分析，是一种广义的线性回归分析模型，常用于数据挖掘，疾病自动诊断，经济预测等领域；探讨引发疾病的危险因素，并根据危险因素预测疾病发生的概率等；针对糖尿病参数进行分析研究，采用LogisticRegressionModels，其定义如下：

$Ln\left[\frac{p}{p-1}\right]=\alpha+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n+e$；估计概率公式为：

$P = 1/[1 + e^{(\alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}]$；其中P表示时间Y发生的概率，即分类结果为糖尿病或非糖尿病的概率；p(Y＝1)＝p/(1-p)表示让步比；Ln[p/(1-p)]是让步比的log值；每一个独立的糖尿病参数X分配相应的系数值β代表该参数对分类结果占有的权重；

在本方法中，最终结果的标签为两类，糖尿病和非糖尿病；数据集中的属性值提供了分类依据；通过LogisticRegression算法分析，得到每一项属性值的权重，从而确定糖尿病参数中的危险因素，进一步分析得到的分类结果即为整体算法模型的预测结果；

在本方法中，使用10折交叉验证方法对预测分析结果进行验证，将初始采样分割成10个子样本，一个单独的子样本被保留作为验证模型的数据，其他9个样本用来训练；交叉验证重复10次，每个子样本验证一次，平均10次的结果或者使用其它结合方式，最终得到一个单一估测；这个方法的优势在于，同时重复运用随机产生的子样本进行训练和验证，每次的结果验证一次；

一种预测过程会有四个不同的结果，分别为TP、TN、FT和FN；在混淆矩阵中显示四种结果相应的数据，TP和TN是分类正确的结果，FT是将原本属于Negative的结果错误分类至Positive类，FN是将原本属于Positive的结果错误分类至Negative类；Precision查准率，是衡量检索系统拒受非相关信息的能力；Recall查全率，是衡量检索系统检出相关信息的能力；MCC，这是一个针对二元分类的有趣性能指标，特别是各个类别在数量上不平衡时；

预警分级模块由C4.5算法和K-means算法组成；

决策树是一个树结构；其每个非叶节点表示一个特征属性上的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶节点存放一个类别；使用决策树进行决策的过程就是从根节点开始，测试待分类项中相应的特征属性，并按照其值选择输出分支，直到到达叶子节点，将叶子节点存放的类别作为决策结果；C4.5是决策树算法的一种，其主要特点是优化信息增益的缺点，提出信息增益率的概念，其定义为$\mathrm{GainRatio(A)} = \frac{Gain(A)}{SplitInfo(A)}$；信息增益率使用“分裂信息”值将信息增益规范化，分裂信息定义如下：$\mathrm{SplitInfo}_A(D) = -\sum_{J=1}^{V} \frac{|D_j|}{|D|} \times \log_2(\frac{|D_j|}{|D|})$，其中$D_1$到$D_V$是V个值的属性A分割D而形成的V个样例子集，分裂信息就是D关于属性A的各值的熵；

选择具有最大增益率的属性作为分裂属性；

在本模块中，首先使用C4.5算法对原始数据集做分类分析，将原始数据集分为糖尿病和非糖尿病两类，分类结果分析得出的分类规则是一系列对属性数值区间的规约，将这些规则应用于下一步的分级定义当中；

K-means算法可以对数据集进行无标签的聚类分析；针对仅包含非糖尿病实例参数的数据集，使用K-means算法进行聚类，将K值设为3，结果生成三个属于不同范围内的类别；通过将结果与现有糖尿病预警分级标准以及上一步生成的若干规则进行对比分析，分别针对高风险、低风险和无风险三个级别得出有实际效用的预警分级规则。

一种基于机器学习的2型糖尿病预测预警方法

技术领域

[0001] 本发明属于机器学习预测分析与医疗健康技术领域，具体涉及一种基于机器学习的2型糖尿病预测预警方法。

背景技术

[0002] 糖尿病是一种以高血糖为特征的慢性疾病，且具有明显的家族遗传特性，接近一半的糖尿病患者有家族遗传病史。国际糖尿病联盟在Diabetes Atlas(Seventh Edition)中的最新数据表明，2015年全世界范围内DM患病人群的数量将近4.15亿。根据近年的增长率预测到2040年，全球糖尿病患者将达到6.42亿，这意味着未来每十个成年人中间就有一个人患有糖尿病。这一惊人的数字毫无疑问需要引起高度重视。

[0003] 近年来，中国已成世界糖尿病患者第一大国，目前患病人数已高达1.1亿人，且患者数量还在不断上升当中。然而我国糖尿病患者知晓率仅为30.1％，其中仅有25.8％的患者得到治疗，而在进行治疗的患者中，血糖得到良好控制的仅有39.7％，据此测算，糖尿病患者中，血糖得到控制的患者比例仅为3.08％。在城市和乡村，上述数据存在显著差异，且不同性别之间差异也较大，经济不发达地区的女性糖尿病的控制情况非常低。在这样的情况下，通过先进的技术手段做好糖尿病患病的前期预防和日常管理就显得格外重要。

[0004] 随着人工智能机器学习等技术的快速发展，大量机器学习算法被运用在医疗健康的方方面面。机器学习是研究如何使用机器来模拟人类学习活动的学科。一种更为严格的定义是：机器学习是一门研究机器获取新知识和新技能，并识别现有知识的学问。机器学习主要研究的是让机器从过去的经历中学习经验，对数据的不确定性进行建模，并在未来进行预测。它是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。

[0005] 糖尿病预测预警是机器学习算法的应用领域之一，主要在如下三个方面展开：1)重要糖尿病参数分析。通过主成分分析方法和关联算法对多种糖尿病数据集的基本属性值进行分析筛选，得出引发糖尿病的重要因素；2)预测模型分析。通过多种分类算法对糖尿病数据集进行有监督的预测分析来判断糖尿病风险参数在一定时间之后引发糖尿病的可能性，主要应用的算法有决策树算法、随机森林算法、神经网络算法以及逻辑回归算法等；3)预警分级分析。针对多种糖尿病数据集，通过无监督的聚类算法，优化现有糖尿病预警分级标准。

发明内容

[0006] 本发明的目的是针对现有糖尿病预测方法准确性的不足，以及对潜在人群进行风险提示的欠缺，提供了一种基于K-means算法、Logistic Regression算法和C4.5算法结合应用的2型糖尿病预测预警方法。本方法对糖尿病参数进行持续的数据收集、数据分析、数据管理等工作，在此基础之上形成预测分析、预警分级和日常管理等功能。

[0007] 为达到上述目的，本发明采用的技术方案为：

[0008] 一种基于机器学习的2型糖尿病预测预警方法，该方法通过K-means算法和Logistic Regression算法建立先聚类再分类的糖尿病双层预测分析模型，对非糖尿病的分类结果通过C4.5算法和K-means算法分析出的规则进行预警分级，对糖尿病数据集通过日常数据收集管理后并进行更新维护以优化原始糖尿病双层预测分析模型。本方法包括数据收集模块、数据预处理模块、预测分析模块、预警分级模块和日常管理模块，数据收集模块与数据预处理模块连接，数据预处理模块与预测分析模块连接，预测分析模块与预警分级模块连接，预警分级模块和日常管理模块连接。

[0009] 本方法包括以下步骤：

[0010] (1)基于现有的健康大数据，获取医院、社康、体检中心等医疗单位内与糖尿病相关的健康数据以建立糖尿病参数数据库，健康数据包括年龄、身高、体重、腰围、臀围、收缩压、舒张压、心率、血糖、血氧、睡眠质量和饮食习惯等。首先对每一个实例以糖尿病和非糖尿病进行标记。

[0011] (2)对糖尿病参数数据库中的原始数据做数据预处理。预处理包括确定统一的数据项以及每一项的具体格式，具体格式为枚举型或数值型等，然后通过数据清洗和数据标准化提高原始数据的质量。

[0012] (3)使用K-means算法和Logistic Regression算法的双层预测分析模型分析处理过的数据集，对每个实例属于糖尿病或是非糖尿病进行预测分类。

[0013] (4)使用C4.5算法和K-means算法结合的分析模型结合现有高危人群划分标准对非糖尿病进行预警分级，提出无风险、低风险和高风险三类标识。

[0014] (5)通过引入最新的糖尿病参数数据，对所有糖尿病参数数据以统一规范的数据格式进行存储并及时反馈至数据收集模块以进行进一步优化。在丰富数据集的基础上，反复步骤(2)、(3)、(4)进行训练以优化本方法的实际应用效果。

[0015] 所述的数据收集模块是基于现有的健康大数据，通过获取医院、社康、体检中心等医疗单位内与糖尿病相关的健康数据以建立糖尿病参数数据库，对每一个实例以糖尿病和非糖尿病进行标记。

[0016] 所述的数据预处理模块分为两个基本步骤，分别是数据清洗和数据标准化。数据清洗包括一致性检查和缺失值处理，一致性检查是根据每个变量的合理取值范围和相互关系，检查数据是否合乎要求，发现超出正常范围、逻辑上不合理或者相互矛盾的数据。缺失值处理即将数据集中缺少的值使用平均值进行替代。数据标准化是在数据清洗完成后，为避免计算过程中数值复杂度，以及避免大数值区间的属性过分支配小数值区间的属性，将所有糖尿病参数中的属性进行Normalize，规范到数值区间[0,1]，使用公式：$\mathrm{Value}=\frac{value-x'}{s}$，其中x'表示糖尿病参数属性的平均值，s表示糖尿病参数属性的标准偏差，Value即糖尿病参数原属性值value进行标准化计算得到的结果。

[0017] 所述的预测分析模块分为两部分，分别是K-means聚类处理和Logistic Regression模型处理。使用K-means聚类算法对剔除了分类标签结果的数据集进行一级处理，设定聚类数目为2，将结果与原始数据集进行对比，剔除聚类错误的数据项，以提供更加准确的数据集进入下一阶段的处理。使用Logistic Regression模型对上述处理过的数据集进行有监督的分类处理，分析结果可得预测准确性。

[0018] 预警分级模块针对糖尿病数据集中的参数，根据C4.5算法和K-means算法基于现有数据分析糖尿病风险层级的规则并与糖尿病分级标准进行对比，以制定更加直接有效的预警分级标准。在此基础上，针对新数据的录入通过新的分级标准进行风险层级判断。

[0019] 所述的日常管理模块包含血糖记录和体征记录。血糖记录针对糖尿病用户，通过引入最新的血糖记录针记录血糖参数数据以更新糖尿病数据库。体征记录针对所有用户，管理的糖尿病参数包括心率、血压、血氧、身高体重和腰臀比等数据。所有参数均以统一规范的数据格式进行存储并及时反馈至数据收集模块进行优化。

[0020] 本发明相对于现有技术，具有以下有益效果：

[0021] 本方法所述数据预处理模块和预测分析模块中使用的混合算法(K-means算法和Logistic Regression算法)在实际应用中相对于现有技术数据处理更清晰、预测准确性更高。针对糖尿病是遗传性疾病的属性，本方法更加专注每个实例的健康参数，包括参数的统一建库管理及更新优化，通过不断引入新的实例数据来优化算法模型的预测准确性。本方法在预测糖尿病与否的基础上，针对非糖尿病实例进行预警分级，能起到更好的防范作用，做到疾病的提早预防。

附图说明

[0022] 图1是本发明方法结构示意图。

[0023] 图2是本发明方法预测分析模块示意图。

[0024] 图3是本发明方法预警分级模块示意图。

[0025] 图4是本发明方法部分数据预测结果示意图。

具体实施方式

[0026] 下面结合附图对本发明作进一步描述。

[0027] 预测分析模块和预警分级模块是本方法的两大核心模块。

[0028] 所述的预测分析模块由K-means算法和Logistic Regression算法组成。

[0029] K-means算法是典型的基于距离的聚类算法，采用距离作为相似度的度量指标，即规定对象间的距离值越小，其相似度越大。K-means算法所产生的簇都是由距离相近的对象组成，故其最终目标是找到这些紧凑且独立的簇。在K-means算法中，K值代表的是初始聚类中心的个数，聚类中心即簇，故K值的选取对聚类结果影响大。

[0030] 针对包含n个糖尿病实例的初始数据集，n为糖尿病实例的个数且取正整数，对应的算法过程如下：

[0031] 1)在给出数据集中的n个糖尿病实例里任意选取2个对象作为初始聚类中心。由于最终的分类结果为两类，将K值定为2；

[0032] 2)对剩余的每个糖尿病实例分别计算与每个中心的距离，并根据剩余的每个糖尿病实例与各个簇中心的距离把剩余的每个糖尿病实例归到最近的中心的簇；

[0033] 3)重新计算每个聚类的中心，判断中心是否发生变化；

[0034] 4)循环步骤2)～3)步直至新的中心与原中心相等或小于指定阈值，即已收敛，则算法结束。通过误差函数判断收敛：$E=\sum_{i=1}^{k}\sum_{x\epsilon C_i} d(x,\mu(C_i))$，其中x表示糖尿病参数中

的每个实例，$\mu(C_i)$表示聚类C_i的中心，$d(x,\mu(C_i))$表示x和$\mu(C_i)$之间的欧几里德距离，k的值为2。

[0035] 在对糖尿病数据集使用K-means算法进行分析时，选取K值为2，将分析结果与数据集原始的分类标签进行比较，剔除错误的噪声数据，将剩余的数据集作为下一级Logistic Regression算法的输入。由于起初K-means算法的Seed值是随机选取的，在剔除数据过程中可能造成错误聚类数量过大，故在每次聚类分析结束后计算数据集剩余比(剩余数据项数量除以原始数据项数量)。若比值大于75%，则进入下一步；若比值小于75%，则进入循环重新选取新的Seed值开始聚类。如附图2所示。

[0036] 以此方法降低人工选取Seed值导致错误分析的风险，并且能够有效控制原始数据集的不必要损失。

[0037] 对高质量的数据集使用Logistic Regression算法进入第二级处理。

[0038] Logistic回归分析，是一种广义的线性回归分析模型，常用于数据挖掘，疾病自动诊断，经济预测等领域。探讨引发疾病的危险因素，并根据危险因素预测疾病发生的概率等。针对糖尿病参数进行分析研究，采用Logistic Regression Models，其定义如下：$\mathrm{Ln}\left[\frac{p}{p-1}\right]=\alpha+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n+e$。估计概率公式为：$P=1/[1+e^{(\alpha+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n)}]$。其中P表示时间Y发生的概率，即分类结果为糖尿病或非糖尿病的概率；$p(Y=1)=p/(1-p)$表示让步比；$Ln[p/(1-p)]$是让步比的log值；每一个独立的糖尿病参数X分配相应的系数值β代表该参数对分类结果占有的权重。

[0039] 在本方法中，最终结果的标签为两类，糖尿病和非糖尿病。数据集中的属性值提供了分类依据。通过Logistic Regression算法分析，得到每一项属性值的权重，从而确定糖尿病参数中的危险因素，进一步分析得到的分类结果即为整体算法模型的预测结果。

[0040] 在本方法中，使用10折交叉验证方法对预测分析结果进行验证，将初始采样分割成10个子样本，一个单独的子样本被保留作为验证模型的数据，其他9个样本用来训练。交叉验证重复10次，每个子样本验证一次，平均10次的结果或者使用其它结合方式，最终得到一个单一估测。这个方法的优势在于，同时重复运用随机产生的子样本进行训练和验证，每次的结果验证一次。

[0041] 图4所示是部分数据预测结果示意图，一种预测过程会有四个不同的结果，分别为True Positive(TP)、True Negative(TN)、False Positive(FT)和False Negative(FN)。在混淆矩阵(Confusion Matrix)中显示四种结果相应的数据，TP和TN是分类正确的结果，FT是将原本属于Negative的结果错误分类至Positive类，FN是将原本属于Positive的结果错误分类至Negative类。Precision查准率，是衡量检索系统拒受非相关信息的能力。Recall查全率，是衡量检索系统检出相关信息的能力。MCC(The Mathews Correlation Coefficient,Mathews相关系数)，这是一个针对二元分类的有趣性能指标，特别是各个类别在数量上不平衡时。

[0042] 预警分级模块由C4.5算法和K-means算法组成。

[0043] 决策树是一个树结构(是二叉树或非二叉树)。其每个非叶节点表示一个特征属性上的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶节点存放一个类别。使用决策树进行决策的过程就是从根节点开始，测试待分类项中相应的特征属性，并按照

其值选择输出分支，直到到达叶子节点，将叶子节点存放的类别作为决策结果。C4.5是决策树算法的一种，其主要特点是优化信息增益的缺点，提出信息增益率的概念，其定义为 $\mathrm{GainRatio(A)} = \frac{Gain(A)}{SplitInfo(A)}$。信息增益率使用“分裂信息”值将信息增益规范化，分裂信息定义如下：$\mathrm{SplitInfo}_A(D) = -\sum_{j=1}^{V} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$ ，其中D_1到D_V是V个值的属性A分割D而形成的V个样例子集，分裂信息就是D关于属性A的各值的熵。

[0044] 选择具有最大增益率的属性作为分裂属性。

[0045] 在本模块中，首先使用C4.5算法对原始数据集做分类分析，将原始数据集分为糖尿病和非糖尿病两类，分类结果分析得出的分类规则是一系列对属性数值区间的规约，将这些规则应用于下一步的分级定义当中。

[0046] K-means算法可以对数据集进行无标签的聚类分析。针对仅包含非糖尿病实例参数的数据集，使用K-means算法进行聚类，将K值设为3，结果生成三个属于不同范围内的类别。通过将结果与现有糖尿病预警分级标准以及上一步生成的若干规则进行对比分析，分别针对高风险、低风险和无风险三个级别得出有实际效用的预警分级规则。

[0047] 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。

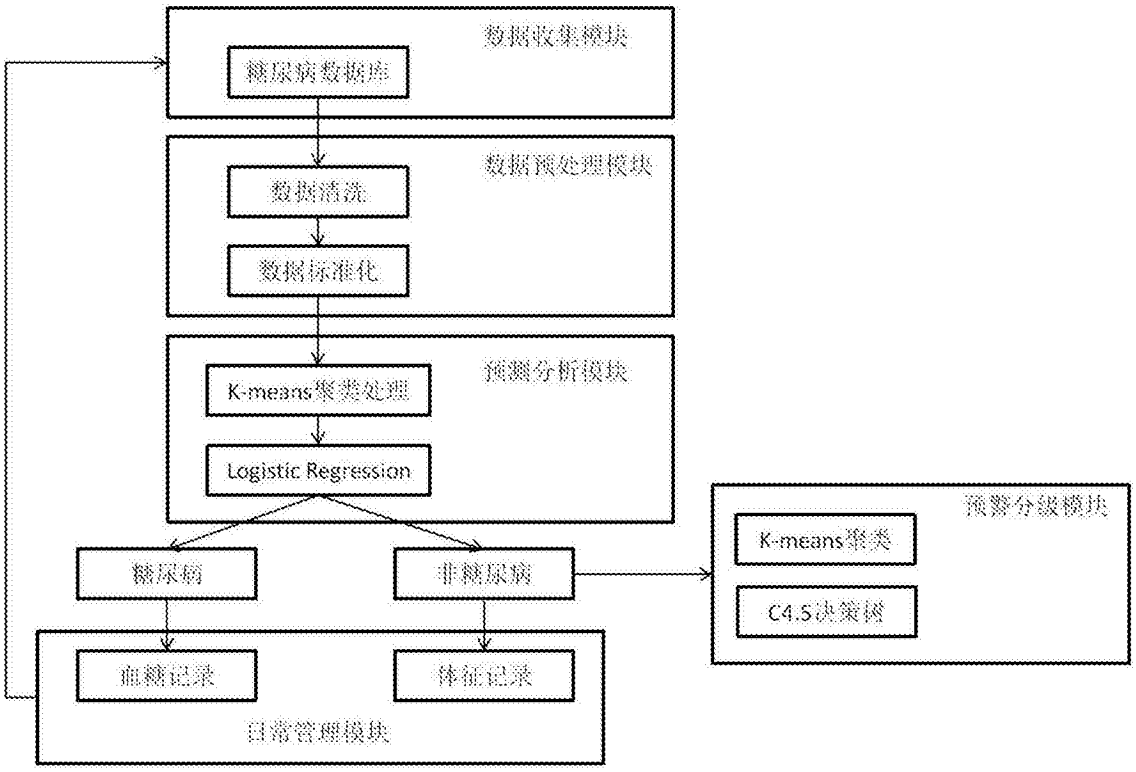

数据收集模块
糖尿病数据库
数据预处理模块
数据清洗
数据标准化
预测分析模块
K-means聚类处理
Logistic Regression
预警分级模块
K-means聚类
C4.5决策树
糖尿病
非糖尿病
血糖记录
体征记录
日常管理模块

图1

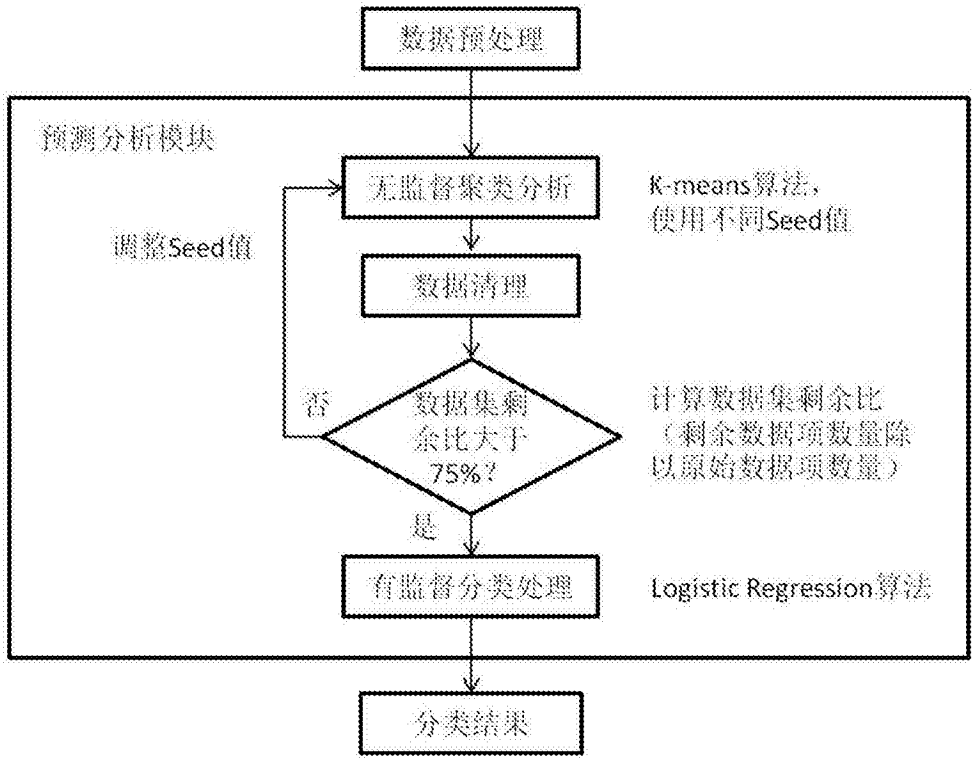

数据预处理
预测分析模块
无监督聚类分析
K-means算法，
使用不同Seed值
调整Seed值
数据清理
否
数据集剩余比大于75%?
计算数据集剩余比（剩余数据项数量除以原始数据项数量）
是
有监督分类处理
Logistic Regression算法
分类结果

图2

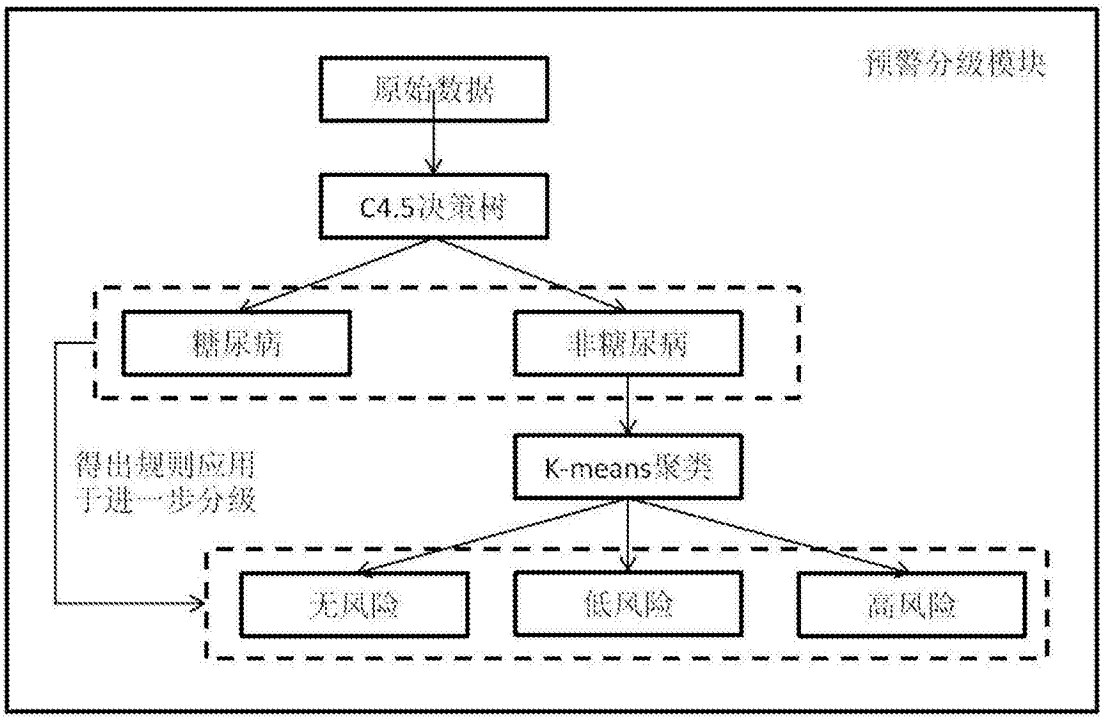


图3

混淆矩阵

	糖尿病	非糖尿病
预测糖尿病	377 (TP)	7 (FP)
预测非糖尿病	20 (FN)	185 (TN)

$$Precision=\frac{TP}{TP+FP} \qquad Recall=\frac{TP}{TP+FN}$$

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}}$$

图4