(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 106909767 B
(45)授权公告日 2021.11.05

(21)申请号 201510964983.5
(22)申请日 2015.12.21
(65)同一申请的已公布的文献号
申请公布号 CN 106909767 A
(43)申请公布日 2017.06.30
(73)专利权人 北京旷博生物技术股份有限公司
地址 100176 北京市大兴区亦庄经济技术开发区地盛东路1号,爱普益大厦2幢3层
(72)发明人 李亦学 张卫红 侯婷 靳文静 王振 孙翔英
(74)专利代理机构 北京尚诚知识产权代理有限公司 11322
代理人 龙淳 顾小曼
(51)Int.Cl.
G16H 50/70(2018.01)
G06K 9/62(2006.01)

(56)对比文件
CN 104232637 A,2014.12.24
CN 102776185 A,2012.11.14
CN 103345544 A,2013.10.09
CN 105139083 A,2015.12.09
CN 105160182 A,2015.12.16
CN 104794321 A,2015.07.22
WO 2015175642 A2,2015.11.19
US 2015066824 A1,2015.03.05
高雪.血浆microRNA作为肝癌及慢性乙肝肝损伤分子标记物的研究.《中国博士学位论文全文数据库 医药卫生科技辑》.2013,(第2期),第E072-31页.

审查员 葛晓倩

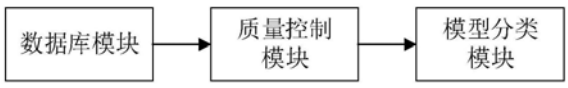
权利要求书1页 说明书10页 附图3页

(54)发明名称

乙肝相关肝硬化分类的系统

(57)摘要

本发明提供了一种用逻辑回归数学模型进行的基于血浆microRNA标志物表达水平的乙肝相关肝硬化分类方法,利用血浆7个microRNA分子标志物组合的表达值,简单准确地诊断乙肝相关肝硬化。其系统的技术方案为:通过收集大量慢性乙型肝炎、乙肝相关肝硬化和健康样本血浆microRNA表达值建立数据库模块,存储用作训练集的原始数据库,及后续盲测数据;建立质量控制模块,去除由于实验误差导致的极端值;建立模型分类模块,通过特征选择等方式构建并优化逻辑回归模型,经评估选择准确率最优的模型建立最终分类方法,采用两层分类模型(健康和肝病(慢性乙型肝炎/肝硬化),慢性乙型肝炎和乙肝相关肝硬化)判定盲测样本分类。



CN 106909767 B

1.一种用于肝病分类的系统,包括数据库模块、质量控制模块、模型分类模块,其中:

所述数据库模块包含作为训练集的原始数据库以及后续收集的盲测数据库;

所述质量控制模块为将由于实验误差导致的极端值去除的模块;

所述模型分类模块包括关于由慢性乙型肝炎/肝硬化和健康对照组成的肝病和健康之间的分类模型DH以及关于慢性乙型肝炎和乙肝相关肝硬化之间的分类模型AB,

所述模型分类模块的建模算法为逻辑回归将多个microRNA分子标志物组合用公式表达,其中区分健康和肝病的DH的算法公式为:

$h_{DH}(x)$ =-1.972X(miR-381-3p)+0.0079X(miR-22-3p)-1.6462X(miR-146a-5p)+74.495

根据最大概率分类可确定的阈值为:

D肝病类:$h_{DH}(x)>0$;

H健康类:$h_{DH}(x)<0$;

其中区分乙肝相关肝硬化和慢性乙型肝炎的AB的算法公式为:

$h_{AB}(x)$ =1.1925X(miR-122-5p)+0.3978X(miR-21-5p)+0.3726X(miR-146a-5p)-1.7062X(miR-29c-3p)+0.1303X(miR-223)+0.8156X(miR-22-3p)-0.1432X $_{ALB}$-0.3608X $_{DNA}$-0.0041X $_{ALT}$-23.9918

A乙肝相关肝硬化类:$h_{AB}(x)>0$,

B慢性乙型肝炎类:$h_{AB}(x)<0$,

其中ALB为白蛋白,DNA为HBV病毒DNA,ALT为转氨酶。

2.根据权利要求1所述的系统,其特征在于,所述的数据库模块中包含10例以上用作训练集的原始数据库以及后续收集的盲测数据库,其中每一例的数据包括miR-122-5p、miR-21-5p、miR-146a-5p、miR-29c-3p、miR-381-3p、miR-223和miR-22-3p相应的Ct值,以及临床指标转氨酶ALT值、白蛋白ALB含量和HBV病毒DNA的含量值。

3.根据权利要求2所述的系统,其特征在于,所述的数据库模块中包含50例以上用作训练集的原始数据库以及后续收集的盲测数据库。

4.根据权利要求2所述的系统,其特征在于,所述的数据库模块中包含200例以上用作训练集的原始数据库以及后续收集的盲测数据库。

5.根据权利要求1所述的系统,其特征在于,所述质量控制模块通过质量控制将由于实验误差导致的极端值去除,非极端Ct值的范围定义为:模型DH中,标志物miR-381-3p的Ct值范围为19.40-32.10,标志物miR-22-3p的Ct值范围为16.72-26.86,标志物miR-146a-5p的Ct值范围为19.32-29.16;模型AB中,标志物miR-122-5p的Ct值范围为17.61-26.99,标志物miR-21-5p的Ct值范围为16.79-24.47,标志物miR-146a-5p的Ct值范围为19.31-26.64,标志物miR-29c-3p的Ct值范围为18.57-26.18,标志物miR-381-3p的Ct值范围为20.13-27.87,标志物miR-223的Ct值范围为15.35-24.15,标志物miR-22-3p的Ct值范围为16.71-23.95。

乙肝相关肝硬化分类的系统

技术领域

[0001] 本发明涉及乙肝相关肝硬化的分类方法和系统，具体来说涉及用逻辑回归数学模型进行基于血浆microRNA标志物表达水平的乙肝相关肝硬化分类的方法和系统。

背景技术

[0002] 我国是病毒性肝炎大国，尤以乙肝患者人数居多。其中乙肝携带者约占总人口的8-10％，其中约有25％发展为慢性乙型肝炎、乙肝相关肝硬化，10％左右发展为肝细胞肝癌(HCC)。乙型肝炎病毒的感染不仅给人民带来了严重的健康危害，而且治疗等疾病相关费用也给患者和国家、社会带来巨大经济负担。

[0003] 目前肝硬化临床诊断手段主要包括组织病理活检、FibroScan、彩色多普勒超声、CT、胃镜、血浆学指标等。但是这些单一技术或指标的临床应用都存在一些局限性和不足，均不能准确、及时诊断肝硬化进展程度，使得对肝硬化的分期诊断仍有赖于肝穿活检病理标准，临床迫切需要一个/一组方便、及时、无创的肝纤维化、肝硬化分级诊断指标。

[0004] microRNA(miRNA)最初发现于1993年，随着高通量测序技术的发展，近年来逐渐成为研究热点。microRNA能够结合于基因序列的侧翼区域阻遏或抑制靶mRNA的翻译，且具有高度的保守性、时序性和组织特异性。近年来的研究表明，肝炎病毒感染、慢性肝炎、肝硬化和microRNA密切相关，microRNA可以通过作用于病毒本身或作用于免疫系统从而影响疾病进程。研究表明，病毒感染的肝病患者microRNA表达谱和健康人组织的microRNA表达谱有明显不同。研究者们还发现在人类血清/血浆中存在大量稳定的小的核糖核酸分子，即microRNA，这为临床上通过检测血清/血浆中microRNA分子表达量诊断肝硬化奠定了基础。

[0005] 综上所述，研究者们虽然已在该领域进行了研究，但是仍面临许多困难和挑战，均未能准确、及时诊断肝硬化进展程度。运用血清/血浆中microRNA标志物表达水平高低，为肝硬化诊断研究提供了新的思路。但目前尚未有关于肝硬化 microRNA标志物或其组合表达变化的深入研究，仍需寻找可有效判断肝硬化的 microRNA标志物或其组合，特别是能将乙肝相关肝硬化与慢性乙型肝炎区别开来的microRNA标志物或其组合，以及基于得到的microRNA标志物组合表达水平，用数学模型构建一种合适且准确的乙肝相关肝硬化分类方法和系统。与传统的肝硬化以及乙型肝炎的诊断方法相比，使用microRNA标志物或其组合的方法具有更快速准确的优点。

发明内容

[0006] 本发明的一个目的是提供了一种用逻辑回归数学模型进行基于血浆 microRNA标志物表达水平的乙肝相关肝硬化分类的方法，包括以下步骤：

[0007] a)使用训练集数据，建立原始数据库；

[0008] b)将上述训练集采用两层分类模型；

[0009] c)通过对上述训练集进行特征选择和数据优化构建并优化所述的逻辑回归数学模型；

[0010] d)进行预测评估；

[0011] e)根据预测评估结果选择最优模型并建立最终的分类方法；

[0012] f)收集独立的测试集样本用于模型的检验和评估。

[0013] 优选地，所述的训练集包含基于血浆microRNA标志物表达的Ct值和临床指标的样本数据；所述的两层分类模型包括关于由慢性乙型肝炎/肝硬化和健康对照组成的肝病和健康分类模型(模型DH)以及关于慢性乙型肝炎和乙肝相关肝硬化分类模型(模型AB)；所述的特征选择采用信息增益算法对训练集特征进行排序来选择贡献度高的特征作为候选microRNA标志物；所述的数据优化的方式为对所述的训练集中的数据进行质量控制和去端值，去掉试验中由于误差导致的极端值，用逻辑回归方法构建所述的逻辑回归数学模型，将多个microRNA分子标志物组合用公式表达：

[0014] $h(x)=h_{\theta}(x)=\theta_0+\theta_1x_1+\theta_2x_2+...+\theta_nx_n$

[0015] 其中$x_1,x_2,...,x_n$是所选取的n个特征，$\theta_0,\theta_1,\theta_2,...\theta_n$是通过训练集得到的各个特征的系数。

[0016] 本发明的第二个目的是提供了一种用于肝病分类的分类的系统，包括数据库模块、质量控制模块、模型分类模块，其中：所述数据库模块包含作为训练集的原始数据库以及后续收集的盲测数据库；所述质量控制模块为将由于实验误差导致的极端值去除的模块；所述模型分类模块包括关于由慢性乙型肝炎/ 肝硬化和健康对照组成的肝病和健康之间的分类模型(模型DH)以及关于慢性乙型肝炎和乙肝相关肝硬化之间的分类模型(肝炎或肝硬化)。所述的肝病为乙肝相关肝病。

[0017] 优选地，所述的数据库模块中包含486例用作训练集的原始数据库以及后续收集的盲测数据，其中每一例的样本包括miR-122-5p、miR-21-5p、miR-146a-5p、 miR-29c-3p、miR-381-3p、

[0018] miR-223和miR-22-3p的Ct表达值，以及临床指标转氨酶(ALT)、白蛋白(ALB) 和HBV病毒DNA的值。

[0019] 所述质量控制模块通过质量控制将由于实验误差导致的极端值去除，所述的非极端值的范围定义为：模型DH中，标志物miR-381-3p的Ct值范围为19.40-32.10，标志物miR-22-3p的Ct值范围为16.72-26.86，标志物miR-146a-5p的Ct值范围为 19.32-29.16；模型AB中，标志物miR-122-5p的Ct值范围为17.61-26.99，标志物 miR-21-5p为16.79-24.47，标志物miR-146a-5p为19.31-26.64，标志物miR-29c-3p 为18.57-26.18，标志物miR-381-3p为20.13-27.87，标志物miR-223为15.35-24.15，标志物miR-22-3p为16.71-23.95。

[0020] 模型分类模块的建模的算法为逻辑回归将多个microRNA分子标志物组合用公式表达，其中区分健康和肝病(DH)的算法公式为：

[0021] $h_{DH}(x)=-1.972X(miR-381-3p)+0.0079X(miR-22-3p)-1.6462X(miR-146a-5p)+74.495$

[0022] 根据最大概率分类可确定的阈值为：

[0023] D肝病(慢性乙型肝炎/肝硬化)类：$h_{DH}(x)>0$；

[0024] H健康类：$h_{DH}(x)<0$；

[0025] 其中区分乙肝相关肝硬化和慢性乙型肝炎(AB)算法公式为：

[0026] $h_{AB}(x)=1.1925X(miR-122-5p)+0.3978X(miR-21-5p)+0.3726X(miR-146a-5p)-$

$1.7062X(miR\text{-}29c\text{-}3p)+0.1303X(miR\text{-}223)+0.8156X(miR\text{-}22\text{-}3p)-0.1432X_{ALB}-0.3608X_{DNA}-0.0041X_{ALT}-23.9918$

[0027] A乙肝相关肝硬化类:$h_{AB}(x)>0$

[0028] B慢性乙型肝炎类:$h_{AB}(x)<0$。

[0029] 本发明的用逻辑回归数学模型进行基于血浆microRNA标志物表达水平的乙肝相关肝硬化分类的方法和系统的优点是提供了一种利用数据库算法和公式,使用microRNA标志物表达Ct值及常见临床指标,自动快速提供乙肝相关肝硬化和慢性乙型肝炎的分类以及结果。

附图说明

[0030] 图1 示出了本发明的用逻辑回归数学模型,进行基于7个血浆microRNA标志物组合表达水平的乙肝相关肝硬化分类方法的建立的实施例的流程图。

[0031] 图2示出了本发明的用逻辑回归数学模型,进行基于7个血浆microRNA标志物组合表达水平的乙肝相关肝硬化分类方法和系统中单层的分类模型的实施例的流程图。

[0032] 图3示例性的示出了本发明的用逻辑回归数学模型,进行基于7个血浆 microRNA标志物组合表达水平的乙肝相关肝硬化分类方法和系统中模型DH特征选择交叉验证结果图。

[0033] 图4示例性的示出了本发明的用逻辑回归数学模型,进行基于7个血浆 microRNA标志物组合表达水平的乙肝相关肝硬化分类方法和系统中模型AB质量控制前后交叉验证结果对比图。图5示出了本发明的用逻辑回归数学模型,进行基于7个血浆microRNA标志物组合表达水平的乙肝相关肝硬化分类方法和系统中实施例的流程图。

具体实施方式

[0034] 本发明通过具体实施例和附图进一步阐述本发明的技术方案,但是本领域普通技术人员可以理解的是:以下具体实施方式以及实施例旨在阐述本发明,而不应理解为以任何方式限制本发明。

[0035] 本发明一个方面是一种新型的用逻辑回归数学模型,进行基于7个血浆 microRNA标志物组合表达水平的乙肝相关肝硬化分类方法。

[0036] 本发明第二个方面提供了用逻辑回归数学模型进行基于血浆microRNA标志物表达水平的乙肝相关肝硬化分类系统。

[0037] 本发明的技术方案是:本发明建立了一种用逻辑回归数学模型进行基于血浆microRNA标志物表达水平的乙肝相关肝硬化分类的方法,包括:

[0038] 收集大量慢性乙型肝炎、乙肝相关乙肝相关肝硬化和健康样本数据并建立原始数据库;

[0039] 采用两层分类模型依次区分健康和慢性乙型肝炎、乙肝相关肝硬化;

[0040] 通过特征选择和数据优化等方式利用训练集构建并优化逻辑回归模型,经过评估后选择最优模型建立最终的分类方法;

[0041] 收集独立的测试集样本用于模型的检验和评估。

[0042] 根据本发明的用逻辑回归数学模型进行基于血浆microRNA标志物表达水平的乙

肝相关肝硬化分类的方法的一实施例，所述的采用两层分类模型，第一层为肝病(慢性乙型肝炎/肝硬化)和健康分类模型(模型DH)，第二层为乙肝相关肝硬化和慢性乙型肝炎分类模型(模型AB)。

[0043] 根据本发明的用逻辑回归数学模型对进行基于血浆microRNA标志物表达水平的进行乙肝相关肝硬化分类的方法的一实施例，所述的通过特征选择和数据优化等方式利用训练集构建并优化逻辑回归模型，特征选择采用信息增益算法对训练集特征进行排序，得到的数据即各个特征的贡献度指标，贡献度高的特征可作为候选microRNA标志物。通过对训练集进行质量控制、单边去端值等方式处理和优化数据，并建立逻辑回归模型，并进行交叉验证评估模型的准确度。将上述过程不断循环得到准确率最佳的模型。

[0044] 上述逻辑回归模型公式为：

[0045] $h(x)=h_{\theta}(x)=\theta_0+\theta_1x_1+\theta_2x_2+...+\theta_nx_n$

[0046] 其中$x_1,x_2,...,x_n$是所选取的n个特征，$\theta_0,\theta_1,\theta_2,...\theta_n$是通过训练集得到的各个特征的系数。

[0047] 本发明还揭示了一种用逻辑回归数学模型进行基于血浆microRNA标志物表达水平的肝病分类的系统，包括数据库模块、质量控制模块、模型分类模块，其中：

[0048] 所述数据库模块，包括用作训练集的原始数据库，以及后续收集的盲测数据；

[0049] 所述质量控制模块，将由于实验误差导致的极端值去除；

[0050] 所述模型分类模块，采用两层分类模型(健康和肝病(慢性乙型肝炎/肝硬化)，乙肝相关肝硬化和慢性乙型肝炎)判定最终的样本分类。

[0051] 根据本发明的用逻辑回归数学模型对进行基于血浆microRNA标志物表达水平的乙肝相关肝硬化分类的方法，所述的数据库模块中存储有包括486例用作训练集的原始数据库，包括150例乙肝相关肝硬化样本，150例慢性乙型肝炎样本，186例健康样本；以及后续收集的盲测数据。每个样本分别包括7个microRNA (miR-122-5p、miR-21-5p、miR-146a-5p、miR-29c-3p、miR-381-3p、miR-223和 miR-22-3p)的表达值，以及三个临床指标的含量值(ALT、ALB和DNA)。

[0052] 根据本发明的用逻辑回归数学模型进行基于血浆microRNA标志物表达水平的乙肝相关肝硬化分类的方法，所述质量控制模块，通过质量控制将由于实验误差导致的极端值去除。非极端值的范围定义为：模型DH中，标志物miR-381-3p 的Ct值范围为19.40-32.10，标志物miR-22-3p的Ct值范围为16.72-26.86，标志物 miR-146a-5p的Ct值范围为19.32-29.16；模型AB中，标志物miR-122-5p的Ct值范围为17.61-26.99，标志物miR-21-5p为16.79-24.47，标志物miR-146a-5p为 19.31-26.64，标志物miR-29c-3p为18.57-26.18，标志物miR-381-3p为20.13-27.87，标志物miR-223为15.35-24.15，标志物miR-22-3p为16.71-23.95。

[0053] 根据本发明的用逻辑回归数学模型对进行基于血浆microRNA标志物表达水平的进行乙肝相关肝硬化分类的方法，所述的模型分类模块，采用两层分类模型 (健康和肝病(慢性乙型肝炎/肝硬化)，乙肝相关肝硬化和慢性乙型肝炎)判定样本分类。建模的算法为逻辑回归，将多个microRNA分子标志物组合用公式表达。其中区分健康和肝病(DH)的算法公式为：

[0054] $h_{DH}(x)=-1.972X(miR-381-3p)+0.0079X(miR-22-3p)-1.6462X(miR-146a-5p)+$

74.495

[0055] 根据最大概率分类可确定的阈值为：

[0056] D肝病(慢性乙型肝炎/肝硬化)类：$h_{DH}(x)>0$；

[0057] H健康类：$h_{DH}(x)<0$；

[0058] 其中区分乙肝相关肝硬化和慢性乙型肝炎(AB)算法公式为：

[0059] $h_{AB}(x)=1.1925X(miR-122-5p)+0.3978X(miR-21-5p)+0.3726X(miR-146a-5p)-1.7062X(miR-29c-3p)+0.1303X(miR-223)+0.8156X(miR-22-3p)-0.1432X_{ALB}-0.3608X_{DNA}-0.0041X_{ALT}-23.9918$

[0060] A乙肝相关肝硬化类：$h_{AB}(x)>0$

[0061] B慢性乙型肝炎类：$h_{AB}(x)<0$。

[0062] 本发明提供了一种用逻辑回归数学模型，进行基于7个血浆microRNA标志物组合表达水平的乙肝相关肝硬化分类方法和系统。本发明的分类方法和系统比传统的临床诊断方法操作简单，且快速。随着大数据时代的到来，测序技术的发展，收集到的健康和疾病的数据不断增加，本发明涉及到的方法会不断改进，得到准确率更高更好的模型。

[0063] 下面结合附图和实施例对本发明作进一步的描述。

[0064] 用逻辑回归数学模型，进行基于7个血浆microRNA标志物组合表达水平的乙肝相关肝硬化分类方法和系统中的实施例

[0065] 图1示出了本发明的用逻辑回归数学模型，进行基于7个血浆microRNA标志物组合表达水平的乙肝相关肝硬化分类方法和系统的实施例的流程。请参见图 1，下面是对本实施例的方法中的各个步骤的详细描述。

[0066] 步骤1：收集大量慢性乙型肝炎、乙肝相关乙肝相关肝硬化和健康样本数据并建立原始数据库。

[0067] 在本步骤中，汇总各大医院收集的慢性乙型肝炎、乙肝相关肝硬化和健康人样本以及相关的临床指标，通过实验提取样本血液测得各样本的microRNA表达值，筛选出差异表达的microRNA分子标志物。样本数据分三类，分别是A乙肝相关肝硬化B慢性乙型肝炎H健康人。

[0068] 步骤2：采用两层分类模型依次区分健康和肝病(慢性乙型肝炎/肝硬化)，乙肝相关肝硬化和慢性乙型肝炎。

[0069] 在本步骤中，模型分两层建立，第一层是模型DH，将乙肝相关肝硬化A和慢性乙型肝炎B归为疾病D，与健康人H分类建模；第二层为模型AB，用于乙肝相关肝硬化A和慢性乙型肝炎B分类建立模型。

[0070] 步骤3：通过特征选择和数据优化等方式利用训练集构建并优化逻辑回归模型，经过评估后选择最优模型建立最终的分类方法；

[0071] 在本步骤中，特征选择的方法为信息增益，对训练集特征进行排序，得到的数据即各个特征的贡献度指标，贡献度高的特征可作为候选microRNA标志物。通过对训练集进行质量控制去掉试验中由于误差导致的极端值。通过单边去端值的方式处理和优化数据，增加不同类别样本之间的区分度。利用处理好的训练集建立逻辑回归模型，并进行交叉验证评估模型的准确度。最后将上述过程不断循环得到最佳的模型。

[0072] 对于逻辑回归模型，计算公式表达为：

[0073] $h(x)=h_{\theta}(x)=\theta_0+\theta_1x_1+\theta_2x_2+...+\theta_nx_n$

[0074] 其中$x_1,x_2,...,x_n$是所选取的n个特征，$\theta_0,\theta_1,\theta_2,...\theta_n$是通过训练集得到的各个特征的系数。n为1＜n≤20的整数，优选小于10。

[0075] 步骤4：收集独立的测试集样本用于模型的检验和评估。

[0076] 在本步骤中，收集独立的测试集样本用于模型的检验和评估，确定模型无过拟合现象。

[0077] 图2示出了本发明的用逻辑回归数学模型，进行基于7个血浆microRNA标志物组合表达水平的乙肝相关肝硬化分类方法和系统中单层的分类模型的实施例的流程图，从图中更清晰地了解单层的分类模型的具体细节。

[0078] 对于上述的步骤，下面是四个具体的实例：

[0079] 实施例1：收集大量慢性乙型肝炎、乙肝相关乙肝相关肝硬化和健康样本数据并建立原始数据库

[0080] 其中，肝硬化诊断依据2000年中华医学会病毒性肝炎防治指南，具体如下：具有肝炎病毒慢性感染病史，影像学提示弥漫肝纤维化，再生结界形成，其他表现可有脾大、脾功能亢进、食管胃底静脉曲张，金标准为病理检查发现再生结节。乙型肝炎诊断依据2000年中华医学会病毒性肝炎防治指南，具体如下：肝炎病程超过半年，或原有乙型肝炎或HBsAg携带史，本次又因同一病原再次出现肝炎症状、体征及肝功能异常，但是没有肝硬化表现的患者，可诊断为慢性乙型肝炎。

[0081] 在2012年6月-2014年3月期间，150个满足以上乙型肝炎定义的患者的血浆样本和150个满足以上乙肝相关肝硬化定义的血浆样本，被预先从首都医科大学附属北京佑安医院采集。经过北京旷博生物技术股份有限公司进行RNA的提取，并完成microRNA的测序分析。将得到的乙肝相关肝硬化和慢性乙型肝炎 microRNA表达值集中构建成数据库，成为原始数据库的部分样本数据集。

[0082] 实施例2：特征选择

[0083] 特征选择的算法为信息增益，是本领域比较成熟的算法，主要借助于weka 下的一个软件包，weka.attributeSelection.InfoGainAttributeEval。可以参考Mitchell, Tom M.(1997).Machine Learning.The Mc-Graw-Hill Companies,Inc.ISBN 0070428077,55页至60页。具体为：将microRNA作为特征，通过weka软件包下InfoGainAttributeEval(信息增益)算法进行特征排序，得到的数据即各个特征的贡献度指标，可作为特征选用的参考。将2015年2月28日测序完成的乙肝相关肝硬化、慢性乙型肝炎和健康样本的多组microRNA数据进行特征选择，对于第一层模型DH来说，随着microRNA分子标志物的增多，模型准确率基本恒定，考虑到标志物数量和模型的准确率，最终选择贡献度排名前三的标志物，分别是miR-381-3p、miR-22-3p和miR-146a-5p，这样，模型的准确率为D 0.997, H 0.986，平均准确率为0.995。请参见图3。

[0084] 实施例3：质量控制

[0085] 将2015年2月28日测序完成的乙肝相关肝硬化、慢性乙型肝炎和健康样本的多组microRNA数据进行质量控制，去掉由于实验误差导致的极端值。极端值定义为：利用R中boxplot软件包做统计，大于最大值或小于最小值为极端值。将模型AB中A和B合起来做初步的质量控制，标志物miR-122-5p的Ct值范围为17.61-26.99，标志物miR-21-5p为16.79-

24.47,标志物miR-146a-5p为19.31-26.64,标志物miR-29c-3p为18.57-26.18,标志物miR-381-3p为20.13-27.87,标志物miR-223 为15.35-24.15,标志物miR-22-3p为16.71-23.95。通过质量控制后,训练集的准确率比质量控制前有所提高,请参见图4。

[0086] 实施例4:最优模型

[0087] 经过多次优化和评估确定最优模型,模型DH的最优模型的特征为3个 microRNA分子标志物(miR-381-3p、miR-22-3p和miR-146a-5p),此时模型交叉验证的准确率为D 0.963,H 0.939,平均准确率为0.954。

[0088] 用逻辑回归数学模型,进行基于7个血浆microRNA标志物组合表达水平的乙肝相关肝硬化分类方法和系统的实施例

[0089] 图5示出了本发明的用逻辑回归数学模型,进行基于7个血浆microRNA标志物组合表达水平的乙肝相关肝硬化分类的系统的组成和连接示意图。请参见图 5。本实施例的系统包括数据库模块、质量控制模块、模型分类模块。

[0090] 数据库模块,即存储包括用作训练集的原始数据库以及后续收集的盲测数据;

[0091] 质量控制模块,是通过质量控制将由于实验误差导致的极端值去除的模块;

[0092] 模型分类模块,是将健康和肝病、乙肝相关肝硬化和慢性乙型肝炎采用两层分类模型进行判定从而进行样本分类的部分。

[0093] 本实施例的系统中数据库模块,存储包括486例用作训练集的原始数据库,包括150例乙肝相关肝硬化样本,150例慢性乙型肝炎样本,186例健康样本;以及后续收集的盲测数据。每个样本包括7个microRNA(miR-122-5p, miR-21-5p,miR-146a-5p,miR-29c-3p,miR-381-3p,miR-223和miR-22-3p)的表达值,以及三个临床指标(ALT、ALB和DNA)。

[0094] 本实施例的系统中质量控制模块,通过质量控制将由于实验误差导致的极端值去除。极端值的范围定义为:模型DH中,标志物miR-381-3p的Ct值范围为 19.40-32.10,标志物miR-22-3p的Ct值范围为16.72-26.86,标志物miR-146a-5p的 Ct值范围为19.32-29.16;模型AB中,标志物miR-122-5p的Ct值范围为17.61-26.99,标志物miR-21-5p为16.79-24.47,标志物miR-146a-5p为19.31-26.64,标志物 miR-29c-3p为18.57-26.18,标志物miR-381-3p为20.13-27.87,标志物miR-223为 15.35-24.15,标志物miR-22-3p为16.71-23.95。

[0095] 本实施例的系统中模型分类模块,采用两层分类模型(健康和肝病,乙肝相关肝硬化和慢性乙型肝炎)判定样本分类。建模的算法为逻辑回归,将多个 microRNA分子标志物组合用公式表达。其中区分健康和肝病(DH)的算法公式为:

[0096] $h_{DH}(x)$ = -1.972X(miR-381-3p)+0.0079X(miR-22-3p)-1.6462X(miR-146a-5p)+74.495

[0097] 根据最大概率分类可确定的阈值为:

[0098] D肝病(慢性乙型肝炎/肝硬化)类:$h_{DH}(x)>0$;

[0099] H健康类:$h_{DH}(x)<0$;

[0100] 其中区分乙肝相关肝硬化和慢性乙型肝炎(AB)算法公式为:

[0101] $h_{AB}(x)$ =1.1925X(miR-122-5p)+0.3978X(miR-21-5p)+0.3726X(miR-146a-5p)-1.7062X(miR-29c-3p)+0.1303X(miR-223)+0.8156X(miR-22-3p)-0.1432X_{ALB} -0.3608X_{DNA}-0.0041X_{ALT}-23.9918

[0102] A乙肝相关肝硬化类:$h_{AB}(x)>0$

[0103] B慢性乙型肝炎类:$h_{AB}(x)<0$。

[0104] 测试实例

[0105] 为了检验本发明的系统的性能,下面使用了两组盲测数据进行验证和评估。

[0106] 盲测数据1

[0107] 盲测数据1是首都医科大学附属北京佑安医院于2014年2月20日完成测序的肝病样本集,包含40例样本,其中乙肝相关肝硬化样本20例,慢性乙型肝炎样本20例。

[0108] 盲测数据2

[0109] 盲测数据2是首都医科大学附属北京佑安医院于2015年4月1日完成测序的肝病和健康样本集,包含40例样本,其中乙肝相关肝硬化样本12例,慢性乙型肝炎样本13例,健康样本15例。

[0110] 系统运行需求/环境

[0111] 1.命令行形式,DOS命令行或者Linux环境下的命令行形式;

[0112] 2.安装有统计软件包R。

[0113] 命令行输入格式:

[0114] RscriptmiRNA.R-itest_DH.txt-typeDH-otest_OH_report.txt-etest_DH_poorQC.txt RscriptmiRNA.R-itest_DH.txt-typeDH-v

[0115] 其中,软件名为miRNA.R,-i输入文件,-type数据处理格式,-o输出文件,-e错误文件,-v直接输出在屏幕上。

[0116] 例1输入文件格式

[0117] sample_name v o e

[0118] 1 27.367422 23.918165 20.387817

[0119] 2 27.591124 24.643553 20.168322

[0120] 3 28.13521 23.20343 21.219599

[0121] 4 27.901966 21.143312 20.402287

[0122] 5 28.58136 20.707237 21.73571

[0123] 6 24.76316 18.762772 19.222338

[0124] 7 27.30698 22.417469 23.841616

[0125] 8 26.368567 19.766613 20.129692

[0126] 9 28.93138 25.612793 21.301153

[0127] 10 26.824923 18.512665 19.730814

[0128] 输出文件格式

[0129] sample_name v o e status

[0130] 1 27.367422 23.918165 20.387817 LiverDisease

[0131] 2 27.591124 24.643553 20.168322 LiverDisease

[0132] 3 28.13521 23.20343 21.219599 LiverDisease

[0133] 4 27.901966 21.143312 20.402287 LiverDisease

[0134] 5 28.58136 20.707237 21.73571 LiverDisease

[0135] 7 27.30698 22.417469 23.841616 LiverDisease

[0136] 8 26.368567 19.766613 20.129692 LiverDisease

[0137] 9 28.93138 25.612793 21.301153 LiverDisease

[0138] 10 26.824923 18.512665 19.730814 LiverDisease

[0139] 结果与讨论

[0140] 盲测数据1

[0141] 盲测数据1只含有肝病数据40例样本，经过质量控制后还剩39例样本，将质控后的数据用于模型AB预测分析评估，详细结果请参见表1，其中，A乙肝相关肝硬化的准确率为0.90，B慢性乙型肝炎的准确率为0.737，平均准确率为 0.821。绘制ROC(receiver operating characteristic)曲线，AUC(曲线下面积，ROC面积)达到0.884。

[0142] 表1模型AB预测分析评估

[0143]

分类	TP 比率	FP 比率	准确率	召回率	F-测量	ROC 面积
A	0.9	0.263	0.783	0.9	0.837	0.884
B	0.737	0.1	0.875	0.737	0.8	0.884
平均	0.821	0.184	0.828	0.821	0.819	0.884

[0144]

[0145] 表1盲测数据1模型AB预测结果

[0146] 盲测数据2

[0147] 盲测数据2含有肝病和健康样本40例，对于第一层模型DH，经过初步质控，保留31例样本全部，将这40例样本数据用于模型DH预测分析评估，详细结果请参见表2，其中D肝病的准确率为0.875，H健康的准确率为1，平均准确率为0.903。绘制ROC曲线，AUC值为0.988。可以明显地看出模型DH的分类效果非常好，也比较符合临床诊断的实际情况。

[0148] 表2

[0149]

分类	TP 比率	FP 比率	准确率	召回率	F-测试	ROC 面积
D	0.875	0	1	0.875	0.933	0.988
H	1	0.125	0.87	1	0.824	0.988
Average	0.903	0.028	0.932	0.903	0.909	0.988

[0150] 表2 盲测数据2模型DH预测结果

[0151] 盲测数据2含有25例乙肝相关肝硬化和慢性乙型肝炎样本，经过质量控制后，还剩19例样本。将质控后的样本用于模型AB预测分析评估，详细结果请参见表3，其中，A乙肝相关肝硬化的准确率为0.90，B慢性乙型肝炎的准确率为0.889，平均准确率为0.895。绘制ROC曲线，AUC值达到0.967。

[0152] 表3

[0153]

分类	TP 比率	FP 比率	准确率	召回率	F-测试	ROC 面积
A	0.9	0.111	0.9	0.9	0.9	0.967
B	0.889	0.1	0.889	0.889	0.889	0.967
Average	0.895	0.106	0.895	0.895	0.895	0.967

[0154] 表3盲测数据2模型AB预测结果

[0155] 综上所述，本发明的系统中模型DH和模型AB的预测分类准确率较高，分类效果很好，没有出现过拟合现象，可以用于实际疾病检测中的乙肝相关肝硬化诊断，且操作简单快速。

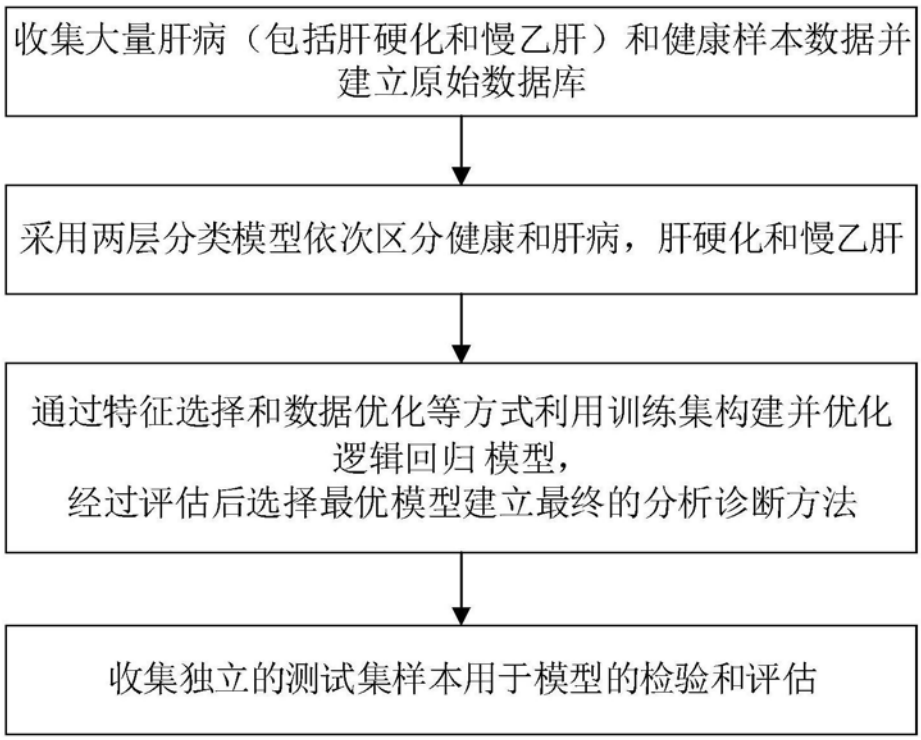

收集大量肝病（包括肝硬化和慢乙肝）和健康样本数据并建立原始数据库
采用两层分类模型依次区分健康和肝病，肝硬化和慢乙肝
通过特征选择和数据优化等方式利用训练集构建并优化逻辑回归 模型，
经过评估后选择最优模型建立最终的分析诊断方法
收集独立的测试集样本用于模型的检验和评估

图1

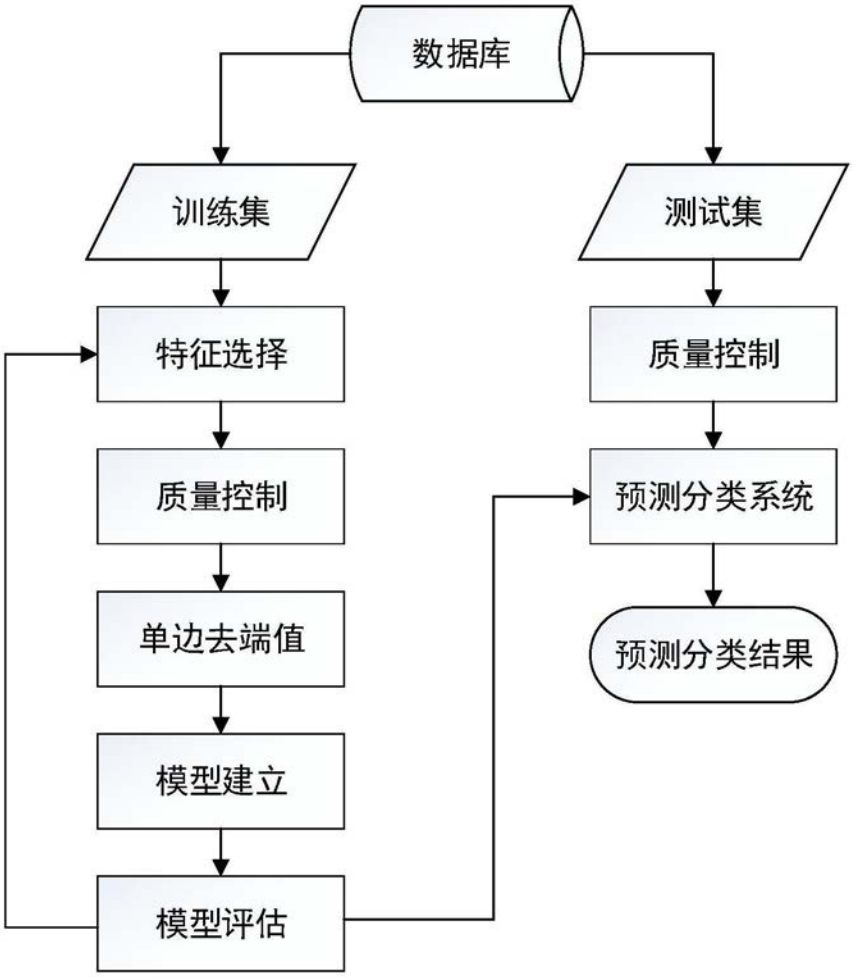


图2

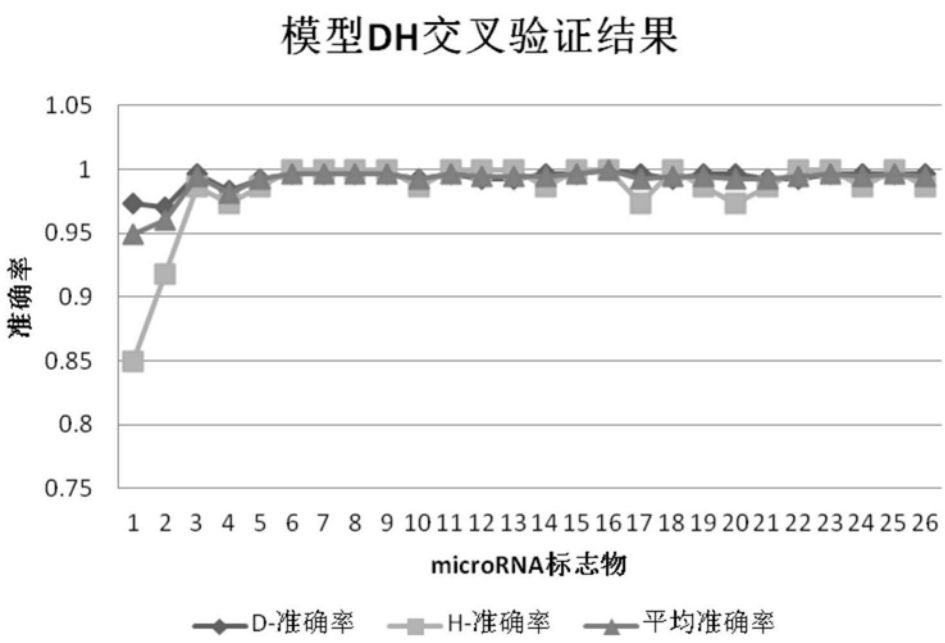

模型DH交叉验证结果
1.05
1
0.95
0.9
0.85
0.8
0.75
准确率
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
microRNA标志物
D-准确率
H-准确率
平均准确率

图3

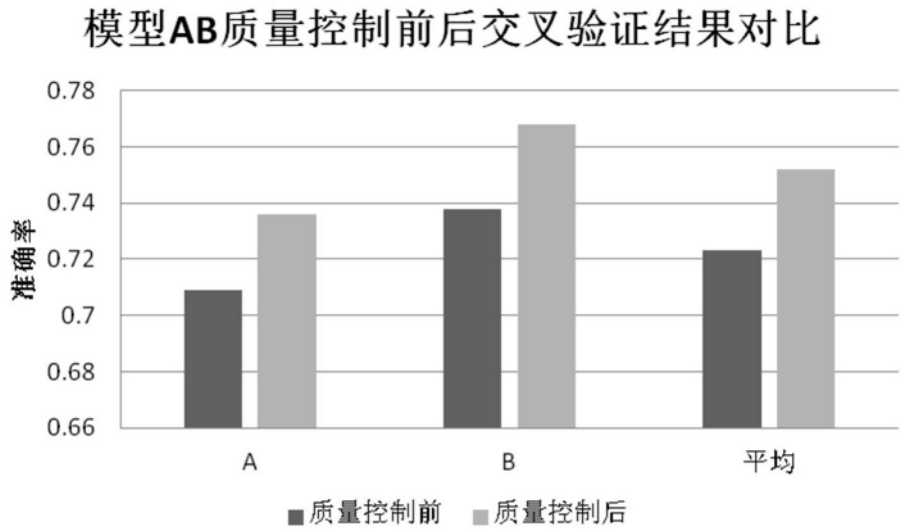

模型AB质量控制前后交叉验证结果对比
0.78
0.76
0.74
0.72
0.7
0.68
0.66
准确率
A
B
平均
质量控制前
质量控制后

图4

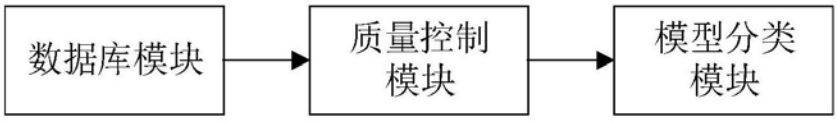

数据库模块
质量控制模块
模型分类模块

图5