

确定染色体非整倍性、构建分类模型的方法和装置

技术领域

5 本发明涉及生物技术领域，特别是无创产前基因检测，具体地涉及确定染色体非整倍性的方法和装置以及相应的构建机器学习分类模型的方法和装置。

背景技术

10 产前筛查方法通常分为两大类，即创伤性方法（也可称之为产前诊断）和非创伤性方法。前者主要包括羊膜腔穿刺、绒毛取样、脐血取样等；后者包括超声波检查、母体外周血清标志物测定和胎儿细胞检测等。通过如绒毛膜绒毛取样（CVS）或羊膜穿刺术等创伤性方法，获得从胎儿处分离得到的细胞，可以利用这些细胞进行常规产前诊断。虽然用这种方法来诊断胎儿非整倍体的准确性较高，但是这些常规的方法是侵入性的，对于孕妇和胎儿都具有一定的危险性。

15 常规的非创伤性筛查方法，例如产前血清学筛查等方法通常准确性较低。
Dennis Lo 等人在母体血浆和血清中发现有非细胞的游离胎儿 DNA，为无创产前筛查（NIPT）提供了新思路。无创性产前筛查主要是利用高通量测序技术分析方法对孕妇外周血中胎儿的游离 DNA 进行分析，以评估胎儿常见染色体非整倍性异常风险。目前筛查范围常见的是 21 号染色体非整倍性（T21）、18 号染色体非整倍性（T18）、13 号染色体非整倍性（T13）和性染色体。

20 基于高通量测序技术，采用孕妇外周血中胎儿游离 DNA 进行胎儿常见染色体非整倍性检测的现有常见技术如下：

1、基于测序序列数定量的方法进行 NIPT：该方法的主要原理是采用比对软件将测序序列（read，有时也称为“测序读段”）定位到预先划定的窗口内，再利用适当的方法对待检染色体进行非整倍性检测。

25 2、基于单核苷酸多态性（SNP）的方法进行 NIPT：该方法的主要原理是依据预先确定的 SNP 位点区域，分别对父母双方基因组 DNA 和胎儿游离 DNA 进行捕获测序，从而利用父母和胎儿的基因型信息采用贝叶斯模型进行待检染色体非整倍性检测。

30 3、基于 DNA 片段大小的方法进行 NIPT：该方法的主要思想是利用双端（PE, paired-end）测序技术，基于胎儿游离 DNA 片段与母体 DNA 片段之间的分布差异特性专门提取胎儿游离 DNA 片段，最后采用 Z 检验基于参照样本对待检染色体进行非整倍性检测。

然而，这些现有的无创产前诊断方法各有相应的缺点，为了方便理解，特总结在下表中：

技术	缺点
基于测序序列数定量的方法进行 NIPT	需要设定阈值，进而引入检测灰区

基于单核苷酸多态性（SNP）的方法进行 NIPT	成本高，检测区域受限于选择的 SNP 位点，需要同时分析提供父母双方单倍型信息
基于 DNA 片段大小的方法进行 NIPT	需要采用 PE 测序，成本较高

因此，目前通过无创手段确定染色体非整倍性的方法仍有待改进。

发明内容

5 本发明旨在至少解决现有技术中存在的技术问题之一。为此，本发明的一个目的在于提出一种能够有效地确定染色体非整倍性的方法。

根据本发明的一个方面，本发明提供了一种确定胎儿是否存在染色体非整倍性的方法，根据本发明的实施例，该方法包括：（1）获取来自孕妇样本的核酸测序数据，所述孕妇样本含有胎儿游离核酸，所述核酸测序数据由多个测序读段构成；（2）基于所述核酸测序数据确定所述孕妇样本的胎儿浓度以及预定染色体的反估浓度，所述反估浓度是基于所述预定染色体的测序读段数目和第一比较染色体的测序读段数目的差异确定的，所述预定染色体包括待测染色体和第二比较染色体，所述第一比较染色体包括至少一个不同于所述预定染色体的常染色体；（3）基于所述待测染色体的反估浓度与所述第二比较染色体的反估浓度的差异确定第一特征，基于所述待测染色体的反估浓度与所述胎儿浓度的差异确定第二特征；和（4）基于所述第一特征和第二特征，利用对照样本的相应数据，确定所述胎儿针对所述待测染色体是否存在非整倍性，其中，所述对照样本包括阳性样本和阴性样本，所述阳性样本针对所述待测染色体具有非整倍性，所述阴性样本针对所述待测染色体不具有非整倍性。

20 通过该方法能够有效地确定胎儿针对待测染色体是否具有非整倍性，另外，根据本发明的实施例，在实施该方法的过程中，发现该方法替代了目前基于测序序列数目中的阈值设定策略，消除了检测灰区，同时还能够缩短样本检测周期，提高客户体验度，并且能够显著降低测序和检测成本。

根据本发明的实施例，上述方法还可以具有下列附加技术特征：

根据本发明的实施例，所述孕妇样本包括孕妇外周血。

25 根据本发明的实施例，所述核酸测序样本是通过双末端测序、单末端测序或者单分子测序获得的。

根据本发明的实施例，所述胎儿浓度是通过下列步骤确定的：（a）将来自所述孕妇样本的所述核酸测序数据与参照序列比对，以便确定落入预定窗口的测序读段的数目；和（b）基于所述落入预定窗口的测序读段的数目，确定所述孕妇样本的胎儿浓度。

30 根据本发明的实施例，在步骤（2）中，所述第一比较染色体的测序读段数目为多条染色体的平均测序读段数目，所述多条常染色体包括至少一个已知不具有非整倍性的常染色体。

根据本发明的实施例，在步骤（2）中，所述第一比较染色体的测序读段数目为至少 15 条常染色体的平均测序读段数目，可选的，第一比较染色体的测序读段数目为至少 20 条常染色体的平均测序读段数目，可选的，第一比较染色体的测序读段数目为全部常染色体的平均测序读段数目。

5 根据本发明的实施例，反估浓度是按照下列公式确定的：

$$F_j = 2 * |R_j - R_r| / (R_r)$$

其中

j 表示需要确定所述反估浓度的染色体的编号，

F_j 表示第 j 号染色体的反估浓度，

10 R_r 表示所述多条常染色体的平均测序读段数目，和

R_j 表示第 j 号染色体的测序读段数目。

根据本发明的实施例，在步骤（3）中，基于所述待测染色体的反估浓度与所述第二比较染色体的反估浓度平均值的差异确定第一特征。

根据本发明的实施例，所述第二比较染色体包含至少 10 条常染色体。

15 根据本发明的实施例，所述第二比较染色体包含 15 条常染色体。

根据本发明的实施例，进一步包括：确定多条常染色体的所述反估浓度；和按照由小至大的优先顺序，选择目标排序的常染色体作为所述第二比较染色体。

根据本发明的实施例，所述第一特征是通过下列公式确定的：

$$X_1 = F_i - F_r$$

20 其中

X_1 表示第一特征，

i 表示所述待测染色体的编号，

F_i 表示所述待测染色体的所述反估浓度，

F_r 表示所述第二比较染色体的反估浓度平均值。

25 根据本发明的实施例，所述第二特征是通过下列公式确定的：

$$X_2 = \frac{F_i}{F_a}$$

其中

X_2 表示第二特征，

i 表示所述待测染色体的编号，

30 F_i 表示所述待测染色体的所述反估浓度，

F_a 表示所述胎儿浓度。

根据本发明的实施例，在进行步骤（4）之前，所述第一特征和所述第二特征进行标准化处理，以便所述第一特征和所述第二特征的绝对值分别独立地处于 0~1 之间。

根据本发明的实施例，在步骤（4）中，所述阳性样本和所述阴性样本的数目比例不低于 1: 4。

根据本发明的实施例，在步骤（4）中，所述阳性样本和所述阴性样本的数目比例不超过 4: 1。

根据本发明的实施例，在步骤（4）中，所述阳性样本和所述阴性样本的数目比例为 1:0.1~5。

5 根据本发明的实施例，在步骤（4）中，所述阳性样本和所述阴性样本的数目比例为 1:0.25~4。

根据本发明的实施例，所述阳性样本和所述阴性样本针对所述待测染色体以外的其他染色体均不存在非整倍性。

根据本发明的实施例，在步骤（4）中，采用所述第一特征和所述第二特征确定所述孕
10 妇样本和所述对照样本的二维特征向量，基于由所述二维特征向量确定的样本间距离，将所述孕妇样本在所述阳性对照样本和所述阴性对照样本之间进行归类，以便确定所述胎儿针对所述待测染色体是否存在非整倍性。

根据本发明的实施例，所述距离为欧几里得距离、曼哈顿距离或切比雪夫距离。

根据本发明的实施例，在步骤（4）中，进一步包括：（4-1）分别计算所述孕妇样本与
15 所述对照样本之间的距离；（4-2）将所得到的所述距离进行排序，所述排序基于由小到大的顺序；（4-3）基于所述排序，从小到大选择预定数量的对照样本；（4-4）分别确定所述预定数量的所述对照样本中阳性样本和阴性样本的数目；（4-5）基于多数决策法，确定将所述孕妇样本的归类结果。

根据本发明的实施例，所述预定数量为不超过 20。

20 根据本发明的实施例，所述预定数量为 3~10。

根据本发明的实施例，在步骤（4-2）中，在进行所述排序之前，预先对所述待测样本与预定所述对照样本之间的距离进行加权处理。

在本发明的第二方面，本发明提供了一种确定胎儿是否存在染色体非整倍性的装置，
25 其特征在于，包括：数据获取模块，用于获取来自孕妇样本的核酸测序数据，所述孕妇样本含有胎儿游离核酸，所述核酸测序数据由多个测序读段构成；胎儿浓度-反估浓度确定模块，用于基于所述核酸测序数据确定所述孕妇样本的胎儿浓度以及预定染色体的反估浓度，所述反估浓度是基于所述预定染色体的测序读段数目和第一比较染色体的测序读段数目的差异确定的，所述预定染色体包括待测染色体和第二比较染色体，所述第一比较染色体包
30 括至少一个不同于所述预定染色体的常染色体；特征确定模块，用于基于所述待测染色体的反估浓度与所述第二比较染色体的反估浓度的差异确定第一特征，基于所述待测染色体的反估浓度与所述胎儿浓度的差异确定第二特征；和非整倍性确定模块，用于基于所述第一特征和第二特征，利用对照样本的相应数据，确定所述孕妇的胎儿针对所述待测染色体是否存在非整倍性，其中，所述对照样本包括阳性样本和阴性样本，所述阳性样本针对所
35 述待测染色体具有非整倍性，所述阴性样本针对所述待测染色体不具有非整倍性。利用根据本发明的实施例的确定胎儿是否存在染色体非整倍性的装置，能够有效地实施前面所描

述的确定胎儿是否存在染色体非整倍性的方法，从而能够有效地确定胎儿针对待测染色体是否存在非整倍性。另外，根据本发明的实施例，在实施该方法的过程中，发现该方法替代了目前基于测序序列数目中的阈值设定策略，消除了检测灰区，同时还能够缩短样本检测周期，提高客户体验度，并且能够显著降低测序和检测成本。

5 根据本发明的实施例，上述装置还可以具有下列附加技术特征：

根据本发明的实施例，所述胎儿浓度-反估浓度确定模块包括：比对单元，用于将来自所述孕妇样本的所述核酸测序数据与参照序列比对，以便确定落入预定窗口的测序读段的数目；和胎儿浓度计算单元，用于基于所述落入预定窗口的测序读段的数目，确定所述孕妇样本的胎儿浓度。

10 根据本发明的实施例，所述胎儿浓度-反估浓度确定模块包括：反估浓度计算单元，用于按照下列公式确定所述反估浓度：

$$F_j = 2 * |R_j - R_r| / (R_r)$$

其中

j 表示需要确定所述反估浓度的染色体的编号，

15 F_j 表示第 j 号染色体的反估浓度，

R_r 表示所述多条常染色体的平均测序读段数目，和

R_j 表示第 j 号染色体的测序读段数目。

根据本发明的实施例，所述胎儿浓度-反估浓度确定模块包括：第二比较染色体确定单元用于将多条常染色体的所述反估浓度按照由小至大的优先顺序，选择目标排序的常染色体作为所述第二比较染色体。

20 根据本发明的实施例，所述特征确定模块包括：

第一特征确定单元，用于通过下列公式确定所述第一特征：

$$X_1 = F_i - F_r$$

其中

25 X_1 表示第一特征，

i 表示所述待测染色体的编号，

F_i 表示所述待测染色体的所述反估浓度，

F_r 表示所述第二比较染色体的反估浓度平均值。

根据本发明的实施例，所述特征确定模块包括：第二特征确定单元，用于通过下列公式确定所述第二特征：

30

$$X_2 = \frac{F_i}{F_a}$$

其中

X_2 表示第二特征，

i 表示所述待测染色体的编号，

35 F_i 表示所述待测染色体的所述反估浓度，

Fa 表示所述胎儿浓度。

根据本发明的实施例，所述特征确定模块包括：标准化处理单元，用于对所述第一特征和所述第二特征进行标准化处理，以便所述第一特征和所述第二特征的绝对值分别独立地处于 0~1 之间。

5 根据本发明的实施例，所述非整倍性确定模块用于采用所述第一特征和所述第二特征确定所述孕妇样本和所述对照样本的二维特征向量，基于由所述二维特征向量确定的样本间距离，将所述孕妇样本在所述阳性对照样本和所述阴性对照样本之间进行归类，以便确定所述胎儿针对所述待测染色体是否存在非整倍性。

根据本发明的实施例，所述距离为欧几里得距离、曼哈顿距离或切比雪夫距离。

10 根据本发明的实施例，所述非整倍性确定模块用于采用 k-近邻模型确定将所述孕妇样本的归类结果。

根据本发明的实施例，所述 k-近邻模型采用的 K 值为不超过 20。

根据本发明的实施例，所述 k-近邻模型采用的 K 值为 3~10。

根据本发明的实施例，所述 k-近邻模型中，对所述样本间距离进行加权处理。

15

在本发明的第三方面，本发明提出了一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现前面所述确定胎儿是否存在染色体非整倍性的方法的步骤。由此，能够有效地实施前面所描述的确定胎儿是否存在染色体非整倍性的方法，从而能够有效地确定胎儿针对待测染色体是否存在非整倍性。另外，根据本发明的
20 实施例，在实施该方法的过程中，发现该方法替代了目前基于测序序列数目中的阈值设定策略，消除了检测灰区，同时还能够缩短样本检测周期，提高客户体验度，并且能够显著降低测序和检测成本。

在本发明的第四方面，本发明提出了一种电子设备，其包括：前面所述的计算机可读
25 存储介质；以及一个或者多个处理器，用于执行所述计算机可读存储介质中的程序。由此，能够有效地实施前面所描述的确定胎儿是否存在染色体非整倍性的方法，从而能够有效地确定胎儿针对待测染色体是否存在非整倍性。另外，根据本发明的实施例，在实施该方法的过程中，发现该方法替代了目前基于测序序列数目中的阈值设定策略，消除了检测灰区，同时还能够缩短样本检测周期，提高客户体验度，并且能够显著降低测序和检测成本。

30

在本发明的第五方面，本发明提出了一种构建机器学习分类模型的方法，根据本发明的实施例，该方法包括：（a）针对多个孕妇样本的每一个分别进行：获取来自所述孕妇样本的核酸测序数据，所述孕妇样本含有胎儿游离核酸，所述核酸测序数据由多个测序读段构成，所述孕妇样本包括至少一个阳性样本和至少一个阴性样本，所述阳性样本针对待测
35 染色体具有非整倍性，所述阴性样本针对所述待测染色体不具有非整倍性；基于所述核酸测序数据确定所述孕妇样本的胎儿浓度以及预定染色体的反估浓度，所述反估浓度是基于

所述预定染色体的测序读段数目和第一比较染色体的测序读段数目的差异确定的，所述预定染色体包括待测染色体和第二比较染色体，所述第一比较染色体包括至少一个不同于所述预定染色体的常染色体；和基于所述待测染色体的反估浓度与所述第二比较染色体的反估浓度的差异确定第一特征，基于所述待测染色体的反估浓度与所述胎儿浓度的差异确定第二特征，（b）将所述多个孕妇样本作为样本，利用所述样本的第一特征和第二特征，进行机器学习训练，以便构建用于确定胎儿是否具有非整倍性的机器学习分类模型。利用该方法，根据本发明的实施例，能够有效地构建机器学习的分类模型，从而进一步可以利用该分类模型对未知的样本进行识别和归类，以确定针对特定的染色体是否存在染色体非整倍性。

10 根据本发明的实施例，所述机器学习分类模型为 KNN 模型。

根据本发明的实施例，所述 KNN 模型采用欧几里得距离。

在本发明的第六方面，本发明提供了一种构建机器学习分类模型的装置，其包括：特征获取模块，用于针对多个孕妇样本的每一个分别进行：获取来自所述孕妇样本的核酸测序数据，所述孕妇样本含有胎儿游离核酸，所述核酸测序数据由多个测序读段构成，所述孕妇样本包括至少一个阳性样本和至少一个阴性样本，所述阳性样本针对待测染色体具有非整倍性，所述阴性样本针对所述待测染色体不具有非整倍性；基于所述核酸测序数据确定所述孕妇样本的胎儿浓度以及预定染色体的反估浓度，所述反估浓度是基于所述预定染色体的测序读段数目和第一比较染色体的测序读段数目的差异确定的，所述预定染色体包括待测染色体和第二比较染色体，所述第一比较染色体包括至少一个不同于所述预定染色体的常染色体；和基于所述待测染色体的反估浓度与所述第二比较染色体的反估浓度的差异确定第一特征，基于所述待测染色体的反估浓度与所述胎儿浓度的差异确定第二特征，训练模块，用于将所述多个孕妇样本作为样本，进行机器学习训练，以便构建用于确定胎儿是否具有非整倍性的机器学习分类模型。利用该装置能够有效地实施前面所述的构建机器学习分类模型的方法，从而能够有效地构建机器学习的分类模型，从而进一步可以利用该分类模型对未知的样本进行识别和归类，以确定针对特定的染色体是否存在染色体非整倍性。

根据本发明的实施例，所述机器学习分类模型为 KNN 模型。

30 在本发明的第七方面，本发明提出了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现权利要求前面所述用于构建机器学习分类方法的步骤。由此，可以有效地实施前面所述的构建机器学习分类模型的方法，从而能够有效地构建机器学习的分类模型，从而进一步可以利用该分类模型对未知的样本进行识别和归类，以确定针对特定的染色体是否存在染色体非整倍性。

35 本发明的附加方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

本发明的上述和/或附加的方面和优点从结合下面附图对实施例的描述中将变得明显和容易理解，其中：

5 图 1 显示了根据本发明一个实施例的确定胎儿是否存在染色体非整倍性的方法的流程示意图；

图 2 显示了根据本发明一个实施例的确定胎儿浓度的方法的流程示意图；

图 3 显示了根据本发明一个实施例的对孕妇样本进行归类的方法的流程示意图；

图 4 显示了根据本发明一个实施例的确定胎儿是否存在染色体非整倍性的装置的框图；

10 图 5 显示了根据本发明一个实施例的胎儿浓度-反估浓度确定模块的框图；

图 6 显示了根据本发明一个实施例的特征确定模块的框图；

图 7 显示了根据本发明一个实施例的构建机器学习分类模型的框图；

图 8 和 9 显示了根据本发明一个实施例利用 KNN 模型对 T21 检测时参数 k 对应的 ROC 曲线；

15 图 10 和 11 显示了根据本发明一个实施例利用 KNN 模型对 T18 检测时参数 k 对应的 ROC 曲线；和

图 12 和 13 显示了根据本发明一个实施例利用 KNN 模型对 T13 检测时参数 k 对应的 ROC 曲线。

20 具体实施方式

下面详细描述本发明的实施例。下面描述的实施例是示例性的，仅用于解释本发明，而不能理解为对本发明的限制。需要说明的是，本申请可用于众多通用或专用的计算装置环境或配置中。例如：个人计算机、服务器计算机、手持设备或便携式设备、平板型设备、多处理器装置、包括以上任何装置或设备的分布式计算环境等等。本申请可以在由计算机执行的计算机可执行指令的一般上下文中描述，例如程序模块。一般地，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。也可以在分布式计算环境中实践本申请，在这些分布式计算环境中，由通过通信网络而被连接的远程处理设备来执行任务。在分布式计算环境中，程序模块可以位于包括存储设备在内的本地和远程计算机存储介质中。

30

根据本发明的一个方面，本发明提供了一种确定胎儿是否存在染色体非整倍性的方法。下面通过参考图 1~3，对根据本发明实施例的确定胎儿是否存在染色体非整倍性的方法进行详细描述。

参考图 1，根据本发明的实施例，该确定胎儿是否存在染色体非整倍性的方法包括：

35 **S100：获取来自孕妇样本的核酸测序数据**

根据本发明的实施例，在该步骤中，首先获取来自孕妇样本的核酸测序数据，该孕妇

样本含有胎儿游离核酸，例如根据本发明的实施例，可以采用的孕妇样本包括但不限于孕妇外周血。如前所述，Dennis Lo 等人在母体血浆和血清中发现有非细胞的游离胎儿 DNA，为无创产前诊断（NIPT）提供了新思路。通过采用孕妇外周血，不会对孕妇造成创伤，避免了由于取样而造成的流产风险。根据本发明的实施例，在获取孕妇样本，例如孕妇外周血，可以对这些样本进行核酸测序，以便获得该孕妇样本的核酸测序数据，通常，该核酸测序数据是由多个或者大量测序读段（read）构成的。根据本发明的实施例，对孕妇样本的核酸分子进行测序的方法并不受特别限制，具体的，可以采用本领域技术人员已知的任何测序方法，例如包括但不限于通过双末端测序、单末端测序或者单分子测序对孕妇样本的核酸分子进行测序。

本领域技术人员能够理解的是，在获得核酸测序数据之后，可以根据质控标准，对所得到的由大量测序读段构成的测序数据进行过滤和筛选处理，除去存在测序质量问题的测序读段，从而可以提高后续数据分析的准确性。

S200 确定反估浓度和胎儿浓度

获取来自孕妇样本的核酸测序数据之后，通过对核酸测序数据的测序读段数目进行分析，可以确定该孕妇样本的胎儿浓度以及特定染色体的反估浓度。

根据本发明的实施例，胎儿浓度是指孕妇样本，例如外周血中的游离核酸中，来自胎儿的游离核酸的数目占总游离核酸数目的比例。通常，该胎儿浓度的数值会随着孕周的增加而提高，例如，在第 12 孕周左右时，胎儿游离核酸（有时直接称为“胎儿游离 DNA”）占总游离核酸的比例（即“胎儿浓度”）可以达到 10~14%，在第 20 孕周之后，这个比例可以达到 20%以上。当胎儿存在异常状况，例如存在染色体非整倍性时，胎儿浓度会出现异常。由此，胎儿浓度可以作为表征孕妇样本的一个重要指标。

本领域技术人员能够通过各种已知的方法获取孕妇样本中的胎儿浓度数据。例如，根据本发明的一个实施例，可以采用包括但不限于 Y 染色体估算法、基于 SNP 的胎儿特异 SNP 位点法、基于核小体印迹法等方法。然而，本发明的发明人发现，这些方法均有其局限性，例如 Y 染色体估算法不适用于女性胎儿，基于 SNP 的胎儿特异 SNP 位点法需要获取父亲的 DNA 样本（有时这些样本是比较难以获得的），基于核小体印迹法的准确性差同时在构建模型时需要进行深度测序。

参考图 2，根据本发明的实施例，可以通过下列步骤确定核酸样本中的胎儿浓度，具体的，包括：

S210：将来自孕妇样本的核酸测序数据与参照序列比对，以便确定落入预定窗口的测序读段的数目；和

S220：基于所述落入预定窗口的测序读段的数目，确定所述孕妇样本的胎儿浓度。

该确定胎儿浓度的方法是基于特定窗口（即一定长度的核酸序列）中的测序读段数目，是与胎儿浓度呈正相关的。因此，通过确定至少一个预定窗口的测序读段的数目，可以反推获得孕妇样本的胎儿浓度，例如加权平均的方式。该预定窗口可以通过统计学的手段或

者机器学习的手段进行确定。根据本发明的实施例，预定窗口是通过对参考基因组序列的特定染色体进行连续划分而获得的，进一步利用各预定窗口的权重，确定胎儿浓度。根据本发明的一些具体示例，各预定窗口的权重是通过利用训练样品预先确定的。由此，结果准确可靠，可重复性好。

5 根据本发明的实施例，权重是利用岭回归统计模型和神经网络模型的至少之一确定的。根据本发明的一些实施例，所述神经网络模型采用 TesnsorFlow 学习系统。根据本发明的一些具体示例，所述 TesnsorFlow 学习系统的参数包括：采用常染色体的各窗口的测序数据数目作为输入层；采用胎儿浓度作为输出层；神经元类型采用 ReLu；优化算法采用选自 Adam、SGD 和 Ftrl 的至少之一；优选 Ftrl。优选地，Tesnsor Flow 学习系统的参数进一步包括：学
10 习速率设置为 0.002；隐藏层的层数为 1；隐藏层中神经元数为 200。由此，结果准确可靠。需要说明的是，本文中所使用的术语“权重”是一个相对的概念，针对某一指标而言。某一指标的权重是指该指标在整体评价中的相对重要程度。例如，某一个“预定窗口的权重”指某一个预订窗口在所有预定窗口中的相对重要程度。某一个“连接权重”指某一个两个不同层连接在所有两个不同层连接中的相对重要程度。

15 关于该确定胎儿浓度的方法，PCT/CN2018/07204（发明名称：确定生物样本中预定来源的游离核酸比例的方法及装置）有详细的介绍，在此不再赘述，在此将该申请的全文以引用的方式全文并入。通过该方法能够简单、快捷、准确地获得胎儿浓度数据，同时，所获得的胎儿浓度数据能够更有效的应用于本发明的方法，用于确定胎儿是否存在染色体非整倍性。

20 另外，获取来自孕妇样本的核酸测序数据之后，不仅可以确定胎儿浓度，还可以进一步确定预定染色体的反估浓度。

在本文中所使用的术语“反估浓度”是指表征特定染色体的 DNA 含量与正常染色体的 DNA 含量之间差异的量度，具体的，可以用特定染色体的测序读段数目与正常染色体的测序读段数目的差异来进行表示。例如，理想状态下，对于存在三体的染色体，其反估浓度为表征多余的一条染色体的 DNA 含量的量，对于正常染色体，则因为没有多余出来的染色体，所以其反估浓度为 0。

因本文主要集中于染色体非整倍性的分析，因此，在本文中所使用的术语“正常染色体”是指不存在染色体非整倍性的染色体，而不意味着该染色体不存在其他的异常状况。

另外，在本文中，多次提到表达方式“.....的测序读段数目”，例如“正常染色体的
30 测序读段数目”，“特定染色体的测序读段数目”，“落入预定对的测序读段数目”等，其含义是指能够与该区域匹配的测序读段数目，例如将核酸测序结果与参考序列例如 hg19 进行比对，例如采用常规软件如 SOAP 等进行比对时，能够与特定的区域比对的测序读段，则被认定为该区域的测序读段。另外，根据本发明的实施例，还可以仅选择“唯一比对测序读段”作为落入特定区域的测序读段，即仅能够与参考序列的一个位置比对上的测序读段。
35 进一步，考虑到测序时有可能存在测序设备受到某些因素的影响造成测序倾向性偏差，例如由于 GC 含量的影响，可以对所得到的测序读段数目进行校正，例如通过 GC 含量进行

校正，具体的，例如，根据本发明的实施例，确定经过校正的测序读段数目的步骤包括：

将参考序列例如人基因组（GRCh37）划分出多个窗口，利用 bwa（0.7.7-r441）把高通量测序后的测序读段与人类参考基因组（GRCh37）进行比对，统计测序读段比对到每条染色体上的每个窗口内的信息，即每个窗口内的测序读段数目，记第 i 个窗口内的测序读段数目为 UR_i ，记参考基因组在第 i 个窗口的 GC 含量为 GC_i 。将各窗口的测序读段数目和 GC 含量进行拟合，并基于拟合系数对原窗口内的测序读段数目进行校正，记第 i 个窗口 GC 校正后的有效序列数为 UR_{Ai} 。

由此，通过选择唯一比对测序读段和进行 GC 含量校正处理，能够有效地提高测序数据分析的准确性和精确性。

如前所述，“反估浓度”是指表征特定染色体的 DNA 含量与正常染色体的 DNA 含量之间差异的量度，因此，该反估浓度可以作为表征孕妇样本的一个重要指标。根据本发明的实施例，反估浓度是基于预定染色体的测序读段数目和第一比较染色体的测序读段数目的差异确定的。

这里所使用的术语“预定染色体”包括待测染色体，即需要确定其是否存在非整倍性的染色体，另外，预定染色体还包括第二比较染色体，根据本发明的实施例，第二比较染色体包括至少一条常染色体。需要说明的是，反估浓度是针对预定染色体的每一条分别进行计算的，因此针对待测染色体和第二比较染色体的每一条，会分别得到与该染色体对应的反估浓度。另外，需要说明的是，对于第一比较染色体和第二比较染色体与待测染色体均来源于相同的样本，而不是采用其他样本的数据进行分析。

根据本发明的实施例，第二比较染色体包含至少 10 条常染色体。根据本发明的实施例，第二比较染色体包含 15 条常染色体。另外，如前所述，反估浓度可以作为表征染色体是否存在异常的一个指标，因此，可以通过借助反估浓度来进行第二比较染色体的选择。根据本发明的实施例，进一步包括：确定多条常染色体的所述反估浓度；和按照由小至大的优先顺序，选择目标排序的常染色体作为所述第二比较染色体。根据前面所描述的，反估浓度越小说明该染色体作为正常染色体的概率越高。例如，通过将所有常染色体按照反估浓度（可以采用绝对反估浓度的绝对值）由小至大进行排序，然后选择反估浓度比较小的排位前 15 的常染色体做第二比较染色体。由此，可以在不确定染色体非整倍性状态的前提下，选择合适的常染色体作为第二比较染色体。当然，本领域技术人员能够理解，在实践中可以通过经验确定其染色体数目是否存在异常的情形，例如通过统计分析发现某些染色体几乎不存在非整倍性，由此，可以将这些染色体作为第二比较染色体。

另外，关于第一比较染色体，如前所述，反估浓度是希望表征特征染色体与正常染色体之间的差异，因此，根据本发明的实施例，第一比较染色体包括至少一个不同于所述预定染色体的常染色体。需要说明的是，这里所说的第一比较染色体和第二比较染色体可能是由交叉的，具体的，在进行反估浓度计算式，会在预定染色体中选择一个特定的染色体，由此，其余的染色体尽管有可能被“第二比较染色体”的含义所覆盖，但仍属于“不同于预定染色体的常染色体”的概念范围。例如，选定第 23 号染色体作为待测染色体，第 2~5

号染色体作为第二比较染色体，则当计算第 23 号染色体的反估浓度时，第 2~5 号染色体仍然可以作为第一比较染色体。另外，根据本发明的实施例，第一比较染色体可以包括多条常染色体，在计算反估浓度时，选择其平均测序读段数目即可。这样，可以进一步提高测序数据分析的效率和准确性。根据本发明的实施例，第一比较染色体的测序读段数目为多条常染色体的平均测序读段数目，该多条常染色体包括至少一个已知不具有非整倍性的常染色体。根据本发明的实施例，第一比较染色体的测序读段数目为至少 15 条常染色体的平均测序读段数目，可选的，第一比较染色体的测序读段数目为至少 20 条常染色体的平均测序读段数目，可选的，第一比较染色体的测序读段数目为全部常染色体的平均测序读段数目。这样，通过选择多条染色体的平均测序读段数目，可以消除各染色体之间的差异。

根据本发明的实施例，反估浓度是按照下列公式确定的：

$$F_j = 2 * |R_j - R_r| / (R_r)$$

其中

j 表示需要确定所述反估浓度的染色体的编号，

F_j 表示第 j 号染色体的反估浓度，

R_r 表示所述多条常染色体的平均测序读段数目，和

R_j 表示第 j 号染色体的测序读段数目。

发明人发现通过该公式计算得到的反估浓度，可以有效地应用于后续的机器学习归类模型。

如前所述，在本步骤中确定的胎儿浓度和反估浓度都受到染色体非整倍性在不同程度上的影响，因此这两个参数可以在后续应用于进行非整倍性的检测。

S300 确定第一特征和第二特征

在确定胎儿浓度和反估浓度之后，可以进一步将这些参数作为样本的特征值，从而可以进一步利用机器学习进行分析。

具体的，根据本发明的实施例，通过待测染色体的反估浓度与第二比较染色体的反估浓度的差异确定第一特征，通过前面所确定的测染色体的反估浓度与胎儿浓度的差异确定第二特征。由此，所得到的第一特征和第二特征，均可以视为能够受到非整倍性影响的特征，因此，可以有效地应用于后续分析。根据本发明的实施例，本领域技术人员可以采用多种算法来表征前面所描述的差异，例如通过计算数值的差，数值的比值等。

如前所述，第二比较染色体的反估浓度优选为多条常染色体的平均反估浓度。由此，可以进一步提高分析的效率和准确性。

另外，根据本发明的实施例，第一特征是通过下列公式确定的：

$$X_1 = F_i - F_r$$

其中

X_1 表示第一特征，

i 表示待测染色体的编号，

Fi 表示待测染色体的所述反估浓度，

Fr 表示第二比较染色体的反估浓度平均值。

根据本发明的实施例，所述第二特征是通过下列公式确定的：

$$X2 = \frac{Fi}{Fa}$$

其中

X2 表示第二特征，

i 表示所述待测染色体的编号，

Fi 表示所述待测染色体的所述反估浓度，

10 Fa 表示所述胎儿浓度。

根据本发明的实施例，如此得到的第一特征和第二特征，一方面均能够体现各自所采用的差异，另一方面所得到的数值均在相同的数量级上，避免出现单个参数过多影响分析结果的情形。如果特征的选择不合适，则后续分析结果有可能会出现偏差，例如

K 模型中要根据样本的特征计算样本之间的距离（例如，样本 x_1 的特征为

15 $(x_1^{(1)}, x_1^{(2)})$ ，样本 x_2 的特征为 $(x_2^{(1)}, x_2^{(2)})$ ，则样本 x_1 和 x_2 之间的距

离为 $d_{12} = \sqrt{(x_1^{(1)} - x_2^{(1)})^2 + (x_1^{(2)} - x_2^{(2)})^2}$ ），若两样本之间特征数值
相 差 特 别 大，例 如 距 离 为

$$d_{12} = \sqrt{(0.23 - 0.15)^2 + (2000 - 3000)^2}$$

20

则尽管二维特征是同等重要的，但显然第二维特征会对距离产生较大的影响。

为了消除这种影响，根据本发明的实施例，在进行后续步骤之前，将所得到的的第一特征和第二特征进行标准化处理，以便第一特征和第二特征的绝对值分别独立地处于 0~1
25 之间。根据本发明的实施例，对第一特征和第二特征进行标准化处理的手段不受特别限制，具体的，可以采用下列方法，针对一批相同维度的数据（均为第一特征或者第二特征），按照下列公式进行处理

$$newValue = (oldVale - min)/(max - min)$$

其中，min 和 max 分别为这批数值的最小和最大值，oldvale 表示处理前的数值，newvalue

表示经过标准化处理后的数值。

由此，可以消除某个特征过多影响最终的分析结果，提高分析结果的准确性。

S400 基于第一特征和第二特征，确定非整倍性

5 如前所述，第一特征和第二特征的数值均受到非整倍性的影响，由此，在获得第一特征和第二特征之后，利用对照样本的相应数据，确定胎儿针对待测染色体是否存在非整倍性。具体的，对照样本包括阳性样本和阴性样本，阳性样本针对待测染色体具有非整倍性，阴性样本针对待测染色体不具有非整倍性。

通过采用第一特征和第二特征作为分类特征，将待测样本针对待测染色体在阳性样本
10 和阴性样本之间进行分类，可以实现所述待测染色体是否存在非整倍性的确定。其中，根据本发明的实施例，发明人在研究过程中，发现，阳性样本和阴性样本的数目满足一定比例能够进一步提高分析的准确性。例如，根据本发明的实施例，阳性样本和阴性样本的数目比例不低于 1:4。根据本发明的实施例，所述阳性样本和所述阴性样本的数目比例不超过 4:1。根据本发明的实施例，所述阳性样本和所述阴性样本的数目比例为 1:0.1~5。根据
15 本发明的实施例，所述阳性样本和所述阴性样本的数目比例为 1:0.25~4。发明人发现，通过采用上述比例可以避免模型结果的偏向性，发明人发现，如果阳性样本偏多时则结果偏阳性，即假阳性率高，反之阴性样本偏多时则结果偏阴性，即假阴性率高。

根据本发明的实施例，所述阳性样本和所述阴性样本针对所述待测染色体以外的其他染色体均不存在非整倍性。由此，可以进一步提高对照样本的归类参考能力。

20 根据本发明的实施例，采用第一特征和第二特征进行归类的方法，并不受特别限制，可以采用多种机器学习的方法，例如神经网络、SVM 法等。发明人在进行深入研究的过程中发现，神经网络需要的训练集的数目比较庞大，而 SVM 则有可能需要额外更多的参数来进行分类，以提高分类的准确性。根据本发明的实施例，可以采用第一特征和第二特征确定所述孕妇样本和所述对照样本的二维特征向量，基于由二维特征向量确定的样本间距离，
25 将孕妇样本在阳性对照样本和阴性对照样本之间进行归类，以便确定胎儿针对所述待测染色体是否存在非整倍性。根据本发明的实施例，，可以采用的距离包括但不限于欧几里得距离、曼哈顿距离或切比雪夫距离。

具体的，根据本发明的实施例，可以采用 K-近邻法（KNN）模型进行归类分析，为了方便理解，参考图 3，对 KNN 模型的过程简单描述如下：

30 根据本发明的实施例，归类处理包括下列步骤：

S410: 分别计算孕妇样本与各个对照样本之间的距离；

S420: 将所得到的距离进行排序，该排序基于由小到大的顺序；

S430: 基于所得到的排序，从小到大选择预定数量的对照样本（这个预定数量即为 KNN 模型中的 K 值）；

35 S440: 分别确定所得到的预定数量的对照样本中阳性样本和阴性样本的数目；

S450: 基于多数决策法，确定将所述孕妇样本的归类结果。

根据本发明的实施例，所述预定数量为不超过 20。根据本发明的实施例，所述预定数量为 3~10。为了方便处理，K 值可以采用奇数，以避免无法做出决定的情形。当然本领域技术人员能够理解的是，对于不同待测染色体最终选择的 K 值可以是不同的，例如，根据本发明的一个实施例，对于 T13 和 T18 检测最终选的 k 为 7，T21 检测最终选择 k 为 9。

5 另外，根据本发明的实施例，在所述排序之前，可以预先对所述待测样本与预定的对照样本之间的距离进行加权处理。由此，可以进一步提高检验的准确性。

本领域技术人员能够理解，这些加权处理的加权系数或者 KNN 模型的 K 值都是可以通过机器学习，利用已知的样本作为训练集进行训练获得的。

具体的，根据本发明的实施例，可以通过下列步骤进行：

10 A、样本集选择

选择有回访结果的样本作为样本集，并按照 6: 2: 2 的比例划分为训练集、测试集和验证集。

B、模型训练

模型的输入： k 值；训练数据集
 15 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, 其中 $x_i \in R^n$ 为样本的 n 维特征向量； $y_i \in \{+1, -1\}, i = 1, 2, \dots, N$ 为样本阴阳性标签（阴性为-1，阳性为+1），N 为样本集大小。

模型的输出：样本 x 所属的类别 y

C：模型验证

20 初始化 $k = 1$ ，基于验证集不断调整 k 值（可采用交叉验证和网格搜索等方法），直到模型的预测能力达到较好的准确性。

D：模型预测

利用训练好的模型对测试集进行预测，从而评估模型的预测性能。

25 由此，通过该方法能够有效地确定胎儿针对待测染色体是否具有非整倍性，另外，根据本发明的实施例，在实施该方法的过程中，发现该方法替代了目前基于测序序列数目中的阈值设定策略，消除了检测灰区，同时还能够缩短样本检测周期，提高客户体验度，并且能够显著降低测序和检测成本。

30 在本发明的第二方面，与上述方法相对应的，本申请实施例还提供了对应的装置，用于实现上述方法。具体的，本发明提供了一种确定胎儿是否存在染色体非整倍性的装置。

参考图 4，该包括确定胎儿是否存在染色体非整倍性的装置包括：

数据获取模块 100，用于获取来自孕妇样本的核酸测序数据，孕妇样本含有胎儿游离核酸，核酸测序数据由多个测序读段构成；

5 胎儿浓度-反估浓度确定模块 200，用于基于核酸测序数据确定孕妇样本的胎儿浓度以及预定染色体的反估浓度，反估浓度是基于预定染色体的测序读段数目和第一比较染色体的测序读段数目的差异确定的，预定染色体包括待测染色体和第二比较染色体，第一比较染色体包括至少一个不同于预定染色体的常染色体；

特征确定模块 300，用于基于待测染色体的反估浓度与第二比较染色体的反估浓度的差异确定第一特征，基于待测染色体的反估浓度与胎儿浓度的差异确定第二特征；和

10 非整倍性确定模块 400，用于基于第一特征和第二特征，利用对照样本的相应数据，确定孕妇的胎儿针对待测染色体是否存在非整倍性，其中，照样本包括阳性样本和阴性样本，所述阳性样本针对所述待测染色体具有非整倍性，所述阴性样本针对所述待测染色体不具有非整倍性。

15 利用根据本发明的实施例的确定胎儿是否存在染色体非整倍性的装置，能够有效地实施前面所描述的确定胎儿是否存在染色体非整倍性的方法，从而能够有效地确定胎儿针对待测染色体是否存在非整倍性。另外，根据本发明的实施例，在实施该方法的过程中，发现该方法替代了目前基于测序序列数目中的阈值设定策略，消除了检测灰区，同时还能够缩短样本检测周期，提高客户体验度，并且能够显著降低测序和检测成本。

参考图 5，根据本发明的实施例，胎儿浓度-反估浓度确定模块 200 包括：

20 比对单元 210，用于将来自孕妇样本的所述核酸测序数据与参照序列比对，以便确定落入预定窗口的测序读段的数目；和

胎儿浓度计算单元 220，用于基于落入预定窗口的测序读段的数目，确定孕妇样本的胎儿浓度。

根据本发明的实施例，胎儿浓度-反估浓度确定模块 200 还包括：

25 反估浓度计算单元 230，用于按照下列公式确定所述反估浓度：

$$F_j = 2 * |R_j - R_r| / (R_r)$$

其中

j 表示需要确定所述反估浓度的染色体的编号，

F_j 表示第 j 号染色体的反估浓度，

30 R_r 表示所述多条常染色体的平均测序读段数目，和

R_j 表示第 j 号染色体的测序读段数目。

根据本发明的实施例，胎儿浓度-反估浓度确定模块 200 包括：

第二比较染色体确定单元 240，用于将多条常染色体的所述反估浓度按照由小至大的优先顺序，选择目标排序的常染色体作为所述第二比较染色体。

35 根据本发明的实施例，特征确定模块 300 包括：

第一特征确定单元 310，用于通过下列公式确定第一特征：

$$X1 = Fi - Fr$$

其中

X1 表示第一特征，

5 i 表示所述待测染色体的编号，

Fi 表示所述待测染色体的所述反估浓度，

Fr 表示所述第二比较染色体的反估浓度平均值。

根据本发明的实施例，特征确定模块 300 还包括：

第二特征确定单元 320，用于通过下列公式确定第二特征：

10
$$X2 = \frac{Fi}{Fa}$$

其中

X2 表示第二特征，

i 表示所述待测染色体的编号，

Fi 表示所述待测染色体的所述反估浓度，

15 Fa 表示所述胎儿浓度。

根据本发明的实施例，特征确定模块 300 还包括：

标准化处理单元 330，用于对所述第一特征和所述第二特征进行标准化处理，以便所述第一特征和所述第二特征的绝对值分别独立地处于 0~1 之间。

20 根据本发明的实施例，非整倍性确定模块 400 用于采用所述第一特征和所述第二特征确定所述孕妇样本和所述对照样本的二维特征向量，基于由所述二维特征向量确定的样本间距离，将所述孕妇样本在所述阳性对照样本和所述阴性对照样本之间进行归类，以便确定所述胎儿针对所述待测染色体是否存在非整倍性。

根据本发明的实施例，所述距离为欧几里得距离、曼哈顿距离或切比雪夫距离。

25 根据本发明的实施例，所述非整倍性确定模块用于采用 k-近邻模型确定将所述孕妇样本的归类结果。

根据本发明的实施例，所述 k-近邻模型采用的 K 值为不超过 20。

根据本发明的实施例，所述 k-近邻模型采用的 K 值为 3~10。

根据本发明的实施例，所述 k-近邻模型中，对所述样本间距离进行加权处理。

30 需要说明的是，前面针对确定胎儿是否存在染色体非整倍性的方法所描述的特征和优点均适用于该确定胎儿是否存在染色体非整倍性的装置，在此不再赘述。

在本发明的第三方面，本发明提出了一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现前面所述确定胎儿是否存在染色体非整倍性的方法的步骤。由此，能够有效地实施前面所描述的确定胎儿是否存在染色体非整倍性的方法，从而能够有效地确定胎儿针对待测染色体是否存在非整倍性。另外，根据本发明的

35

实施例，在实施该方法的过程中，发现该方法替代了目前基于测序序列数目中的阈值设定策略，消除了检测灰区，同时还能够缩短样本检测周期，提高客户体验度，并且能够显著降低测序和检测成本。

本领域技术人员能够理解的是，前面针对确定胎儿是否存在染色体非整倍性的方法所描述的特征和优点均适用于该计算机可读存储介质，在此不再赘述。

在本发明的第四方面，本发明提出了一种电子设备，其包括：前面所述的计算机可读存储介质；以及一个或者多个处理器，用于执行所述计算机可读存储介质中的程序。由此，能够有效地实施前面所描述的确定胎儿是否存在染色体非整倍性的方法，从而能够有效地确定胎儿针对待测染色体是否存在非整倍性。另外，根据本发明的实施例，在实施该方法的过程中，发现该方法替代了目前基于测序序列数目中的阈值设定策略，消除了检测灰区，同时还能够缩短样本检测周期，提高客户体验度，并且能够显著降低测序和检测成本。本领域技术人员能够理解的是，前面针对确定胎儿是否存在染色体非整倍性的方法所描述的特征和优点均适用于该电子设备，在此不再赘述。

在本发明的第五方面，本发明提出了一种构建机器学习分类模型的方法，根据本发明的实施例，该方法包括：

(a) 针对多个孕妇样本的每一个分别进行：

获取来自孕妇样本的核酸测序数据，孕妇样本含有胎儿游离核酸，核酸测序数据由多个测序读段构成，孕妇样本包括至少一个阳性样本和至少一个阴性样本，阳性样本针对待测染色体具有非整倍性，阴性样本针对待测染色体不具有非整倍性；

基于核酸测序数据确定孕妇样本的胎儿浓度以及预定染色体的反估浓度，反估浓度是基于预定染色体的测序读段数目和第一比较染色体的测序读段数目的差异确定的，预定染色体包括待测染色体和第二比较染色体，第一比较染色体包括至少一个不同于预定染色体的常染色体；和基于待测染色体的反估浓度与第二比较染色体的反估浓度的差异确定第一特征，基于待测染色体的反估浓度与胎儿浓度的差异确定第二特征，

(b) 将多个孕妇样本作为样本，利用样本的第一特征和第二特征，进行机器学习训练，以便构建用于确定胎儿是否具有非整倍性的机器学习分类模型。

利用该方法，根据本发明的实施例，能够有效地构建机器学习的分类模型，从而进一步可以利用该分类模型对未知的样本进行识别和归类，以确定针对特定的染色体是否存在染色体非整倍性。根据本发明的实施例，机器学习分类模型为 KNN 模型。根据本发明的实施例，KNN 模型采用欧几里得距离。

本领域技术人员能够理解的是，前面针对确定胎儿是否存在染色体非整倍性的方法所描述的特征和优点均适用于该构建模型的方法，在此不再赘述。

在本发明的第六方面，本发明提供了一种构建机器学习分类模型的装置。

参考图 7，该装置包括：

特征获取模块 800，用于针对多个孕妇样本的每一个分别进行：获取来自孕妇样本的核酸测序数据，孕妇样本含有胎儿游离核酸，核酸测序数据由多个测序读段构成，孕妇样本包括至少一个阳性样本和至少一个阴性样本，阳性样本针对待测染色体具有非整倍性，阴性样本针对待测染色体不具有非整倍性；基于核酸测序数据确定孕妇样本的胎儿浓度以及预定染色体的反估浓度，反估浓度是基于预定染色体的测序读段数目和第一比较染色体的测序读段数目的差异确定的，预定染色体包括待测染色体和第二比较染色体，第一比较染色体包括至少一个不同于预定染色体的常染色体；和基于待测染色体的反估浓度与第二比较染色体的反估浓度的差异确定第一特征，基于待测染色体的反估浓度与胎儿浓度的差异确定第二特征；和

训练模块 900，用于将多个孕妇样本作为样本，进行机器学习训练，以便构建用于确定胎儿是否具有非整倍性的机器学习分类模型。利用该装置能够有效地实施前面的构建机器学习分类模型的方法，从而能够有效地构建机器学习的分类模型，从而进一步可以利用该分类模型对未知的样本进行识别和归类，以确定针对特定的染色体是否存在染色体非整倍性。

根据本发明的实施例，机器学习分类模型为 KNN 模型。

利用该装置，根据本发明的实施例，能够有效地构建机器学习的分类模型，从而进一步可以利用该分类模型对未知的样本进行识别和归类，以确定针对特定的染色体是否存在染色体非整倍性。根据本发明的实施例，机器学习分类模型为 KNN 模型。根据本发明的实施例，KNN 模型采用欧几里得距离。

本领域技术人员能够理解的是，前面针对确定胎儿是否存在染色体非整倍性的方法所描述的特征和优点均适用于该构建模型的装置，在此不再赘述。

在本发明的第七方面，本发明提出了一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现权利要求前面用于构建机器学习分类方法的步骤。由此，可以有效地实施前面的构建机器学习分类模型的方法，从而能够有效地构建机器学习的分类模型，从而进一步可以利用该分类模型对未知的样本进行识别和归类，以确定针对特定的染色体是否存在染色体非整倍性。本领域技术人员能够理解的是，前面针对确定胎儿是否存在染色体非整倍性的方法所描述的特征和优点均适用于该构建模型的计算机可读存储介质，在此不再赘述。

下面将结合实施例对本发明的方案进行解释。本领域技术人员将会理解，下面的实施例仅用于说明本发明，而不应视为限定本发明的范围。实施例中未注明具体技术或条件的，按照本领域内的文献所描述的技术或条件或者按照产品说明书进行。实施例中未注明具体条件者，按照常规条件或制造商建议的条件进行。所用试剂或仪器未注明生产厂商者，均为可以通过市场获得的常规产品。

实施例 1：

本实施例基于 BGISEQ-500 平台 2017 年到 2018 年间的有回访结果的 3075 例样本（其中男胎：1716 例，女胎：1359 例，阴性样本：2215 例，21 号染色体三体（T21）：637 例，18 号染色体三体（T18）：165 例，13 号染色体三体（T13）：58 例）进行模型训练和模型预测。

- 5 首先，将参考基因组(GRCh37)按固定长度(本方法使用 60K)连续划分相邻的窗口，过滤掉 N 区内的窗口，统计窗口内 GC 含量，得到参照窗口文件 hg19.gc；

接下来，将基于 CG 平台 SE 测序之后的序列（35bp）比对（BWA V0.7.7-r441）到参考基因组(GRCh37)；

- 10 过滤及初步统计：根据比对结果选择唯一完全比对的序列，去掉重复序列和存在碱基错配的序列得到有效序列，然后按照 hg19.gc 文件中窗口统计各个窗口的有效序列数和其 GC 含量；

GC 矫正，步骤具体如下：

- 15 对于某个样本，记第 i 个窗口的有效序列数为 UR_i ，记参考基因组在该窗口的 GC 含量为 GC_i （hg19.gc 文件中记录），记常染色体(1~22 号染色体)上所有窗口有效序列数均值为 $\overline{UR}$ ；

利用常染色体所有窗口的有效序列数及 GC 含量进行拟合(本实施例中使用三次样条拟合)得到二者之间的关系式： $ur = f(gc)$ ；

对所有染色体的窗口进行校正： $UR_i * (\overline{UR} / f(GC_i))$ ， $i = 1, 2, 3, \dots, N$ ；

记第 i 个窗口 GC 校正后的有效序列数为 UR_{A_i} 。

- 20 按照下列公式计算各染色体的反估浓度：

反估浓度的计算公式如下：

$$F_j = \begin{cases} 2 \times \frac{\overline{URAA}_j - \overline{URAA}}{\overline{URAA}}, & j = 1, \dots, 22 \\ 2 \times \frac{\overline{URAA} - \overline{URAA}_j}{\overline{URAA}}, & j = X \end{cases}$$

j 表示染色体的编号， $\overline{URAA}_j$ 表示能够与第 j 号染色体参照序列匹配的经过 GC 校正后的测

- 25 序读段数目， $\overline{URAA}$ 表示能够与所有常染色体参照序列匹配的经过 GC 校正的平均测序读段数目。

按照常规方法或者 PCT/CN2018/072045 所公布的方法确定胎儿浓度。

基于样本集进行 KNN 模型训练和样本预测，具体步骤如下：

(a) 样本集划分和数据预处理：将样本集以 6:2:2 的比例随机分为训练集、验证集和测试集；分别对训练集、验证集和测试集的样本进行数据预处理，使每个样本得到一个二维的特征向量，以及对应的标签（阴性为-1，阳性为+1）。

5 (b) 超参数 k 的选择：发明人发现如果选择较小的 k 值，就相当于用较小的邻域中的训练样本集进行预测，预测结果会对邻近的样本点非常敏感，整体模型变得复杂，容易发生拟合；如果选择较大的 k 值，就相当于用较大邻域中的训练样本集进行预测，这时与新输入样本较远（不相似）的训练样本集也会对预测结果起作用，使预测发生错误；一种极限情况是 K 为某个数值时，无论新输入的样本属于什么类别，都将简单的预测为在
10 训练样本集中最多的类。因此，在本发明的实践中， k 一般取一个比较小的数值。

(c) 模型训练：包括两部分：KNN 模型训练和 k 值的选择。此时选取欧氏距离和多数表决规则。

KNN 模型训练：对于分类决策函数：

$$f:R^n \rightarrow \{c_1, c_2\} = \{-1, +1\}$$

15 其中 $x \in R^n$ 为 n 维特征空间，-1 和+1 分别为样本标签（阴性为-1，阳性为+1）。

那么误分类的概率为：

$$P(Y \neq f(X)) = 1 - P(Y = f(X))$$

对与给定的样本 $x \in X$ ，其最近邻的 k 个训练样本点构成的集合为 $N_k(x)$ 。如果涵盖 $N_k(x)$ 的区域类别为 c_j ，那么误分类的概率为：

$$\frac{1}{k} \sum_{x_i \in N_k(x)} I(y_i \neq c_j) = 1 - \frac{1}{k} \sum_{x_i \in N_k(x)} I(y_i = c_j)$$

20

要使误分类概率最小，就要使 $\frac{1}{k} \sum_{x_i \in N_k(x)} I(y_i = c_j)$ 最大。因此选定 k 值

$$\frac{1}{k} \sum_{x_i \in N_k(x)} I(y_i = c_j)$$

后，模型训练的过程即为使最大的过程。

k 值的筛选：初始化 $k = 1$ ($k \in \{1, 2, \dots, 20\}$)，基于验证集采用线性搜索的方法确定 k 值。结果见图 8~13，其中，图 8~13 均是 ROC 曲线图，分别表示参数 k 选择不同数值时对应的 ROC 曲线图，其反映对应的分类器的效果，评价标准就是 AUC，即 ROC 曲线下面的面积，AUC 越大，分类性能越好。图 8 和 9 表示 KNN 模型对 T21 检测时，参数 k 分别选择 6,7,8 和 9 时的 ROC 曲线图。图 10 和 11 表示 KNN 模型对 T18 检测时，参数 k 分别选择 6,7,8 和 9 时的 ROC 曲线图。图 12 和 13 表示 KNN 模型对 T13 检测时，参数 k 分别选择 6,7,8 和 9 时的 ROC 曲线图。根据图 8~13 的结果，对于 T13 和 T18 最终选的 k 为 7，T21 最终选择 k 为 9。

(d) 模型预测：基于上述步骤训练好的模型对测试集进行预测，预测结果如下表所示。

T21		核型结果	
		阳性样本量	阴性样本量
KNN 模型结果	阳性样本量	122	3
	阴性样本量	0	483

T18		核型结果	
		阳性样本量	阴性样本量
KNN 模型结果	阳性样本量	33	5
	阴性样本量	0	570

T13		核型结果	
		阳性样本量	阴性样本量
KNN 模型结果	阳性样本量	10	6
	阴性样本量	0	592

分别计算检测的灵敏度、特异度、PPV 和 ACC 的结果如下表所示。

	灵敏度	特异度	PPV	ACC
T21	100%	99.38%	97.60%	99.51%

T18	100%	99.13%	86.84%	99.18%
T13	100%	99.00%	62.50%	99.01%

2.5 与 SVM 模型的比较

基于相同的训练集、验证集和测试集，用 SVM（支持向量机）的方法对样本的阴阳性进行分类，结果如下：

T21		核型结果	
		阳性样本量	阴性样本量
SVM 模型结果	阳性样本量	122	14
	阴性样本量	0	472

5

T18		核型结果	
		阳性样本量	阴性样本量
SVM 模型结果	阳性样本量	33	8
	阴性样本量	0	567

T13		核型结果	
		阳性样本量	阴性样本量
SVM 模型结果	阳性样本量	10	8
	阴性样本量	0	590

分别计算检测的灵敏度、特异度、PPV 和 ACC 的结果如下表所示

	灵敏度	特异度	PPV	ACC
T21	100%	97.13%	89.71%	97.71%
T18	100%	98.61%	80.49%	98.69%
T13	100%	98.67%	55.56%	98.69%

- 10
- 从数据中可以看出，不管 KNN 模型还是 SVM 模型，在测试集中，T13、T18 和 T21 的检测均没有漏检，灵敏度都达到了 100%。但是在 T21 的检测中，SVM 模型有 14 例假阳性样本，而 KNN 模型只有 3 例假阳；在 T18 的检测中，SVM 模型有 8 例假阳，而 KNN 模型只有 5 例假阳；在 T13 的检测中，SVM 模型有 8 例假阳，而 KNN 模型有 6 例假阳。不管对 T21、T18 还是 T13，KNN 模型均比 SVM 模型的假阳率低。
- 15
- 发明人分析，KNN 模型较 SVM 模型假阳性率低的主要原因是：模型本身导致，即 KNN 主要依据聚类，它是很多个细化的聚类，而 SVM 只是简单的两类，所以细致程度上没有 KNN 好。

在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

5

尽管已经示出和描述了本发明的实施例，本领域的普通技术人员可以理解：在不脱离本发明的原理和宗旨的情况下可以对这些实施例进行多种变化、修改、替换和变型，本发明的范围由权利要求及其等同物限定。

权利要求书

1、一种确定胎儿是否存在染色体非整倍性的方法，其特征在于，包括：

5 (1) 获取来自孕妇样本的核酸测序数据，所述孕妇样本含有胎儿游离核酸，所述核酸测序数据由多个测序读段构成；

(2) 基于所述核酸测序数据确定所述孕妇样本的胎儿浓度以及预定染色体的反估浓度，所述反估浓度是基于所述预定染色体的测序读段数目和第一比较染色体的测序读段数目的差异确定的，所述预定染色体包括待测染色体和第二比较染色体，所述第一比较染色体包括至少一个不同于所述预定染色体的常染色体；

10 (3) 基于所述待测染色体的反估浓度与所述第二比较染色体的反估浓度的差异确定第一特征，基于所述待测染色体的反估浓度与所述胎儿浓度的差异确定第二特征；和

(4) 基于所述第一特征和第二特征，利用对照样本的相应数据，确定所述胎儿针对所述待测染色体是否存在非整倍性，其中，所述对照样本包括阳性样本和阴性样本，所述阳性样本针对所述待测染色体具有非整倍性，所述阴性样本针对所述待测染色体不具有非整倍性。

15 2、根据权利要求 1 所述的方法，其特征在于，所述孕妇样本包括孕妇外周血。

3、根据权利要求 1 所述的方法，其特征在于，所述核酸测序样本是通过双末端测序、单末端测序或者单分子测序获得的。

4、根据权利要求 1 所述的方法，其特征在于，所述胎儿浓度是通过下列步骤确定的：

20 (a) 将来自所述孕妇样本的所述核酸测序数据与参照序列比对，以便确定落入预定窗口的测序读段的数目；和

(b) 基于所述落入预定窗口的测序读段的数目，确定所述孕妇样本的胎儿浓度。

5、根据权利要求 1 所述的方法，其特征在于，在步骤 (2) 中，所述第一比较染色体的测序读段数目为多条常染色体的平均测序读段数目，所述多条常染色体包括至少一个已知不具有非整倍性的常染色体。

6、根据权利要求 5 所述的方法，其特征在于，在步骤 (2) 中，所述第一比较染色体的测序读段数目为至少 15 条常染色体的平均测序读段数目，

可选的，第一比较染色体的测序读段数目为至少 20 条常染色体的平均测序读段数目，

可选的，第一比较染色体的测序读段数目为全部常染色体的平均测序读段数目。

30 7、根据权利要求 5 所述的方法，其特征在于，反估浓度是按照下列公式确定的：

$$F_j = 2 * |R_j - R_r| / (R_r)$$

其中

j 表示需要确定所述反估浓度的染色体的编号，

F_j 表示第 j 号染色体的反估浓度，

35 R_r 表示所述多条常染色体的平均测序读段数目，

R_j 表示第 j 号染色体的测序读段数目。

8、根据权利要求 1 所述的方法，其特征在于，在步骤（2）中，所述第二比较染色体包含多个不具有非整倍性的常染色体，并且在步骤（3）中，基于所述待测染色体的反估浓度与所述第二比较染色体的反估浓度平均值的差异确定第一特征。

5 9、根据权利要求 8 所述的方法，其特征在于，所述第二比较染色体包含至少 10 条常染色体。

10、根据权利要求 8 所述的方法，其特征在于，所述第二比较染色体包含 15 条常染色体。

11、根据权利要求 8 所述的方法，其特征在于，进一步包括：

10 确定多条常染色体的所述反估浓度；和

按照由小至大的优先顺序，选择目标排序的常染色体作为所述第二比较染色体。

12、根据权利要求 1 所述的方法，其特征在于，所述第一特征是通过下列公式确定的：

$$X1 = Fi - Fr$$

其中

15 $X1$ 表示第一特征，

i 表示所述待测染色体的编号，

Fi 表示所述待测染色体的反估浓度，

Fr 表示所述第二比较染色体的反估浓度平均值。

13、根据权利要求 12 所述的方法，其特征在于，所述第二特征是通过下列公式确定的：

20
$$X2 = \frac{Fi}{Fa}$$

其中，

$X2$ 表示第二特征，

i 表示所述待测染色体的编号，

Fi 表示所述待测染色体的反估浓度，

25 Fa 表示所述胎儿浓度。

14、根据权利要求 1~13 任一项所述的方法，其特征在于，在进行步骤（4）之前，对所述第一特征和所述第二特征进行标准化处理，以便所述第一特征和所述第二特征的绝对值分别独立地处于 0~1 之间。

15、根据权利要求 1 所述的方法，其特征在于，在步骤（4）中，所述阳性样本和所述
30 阴性样本的数目比例不低于 1: 4。

16、根据权利要求 1 所述的方法，其特征在于，在步骤（4）中，所述阳性样本和所述阴性样本的数目比例不超过 4: 1。

17、根据权利要求 1 所述的方法，其特征在于，在步骤（4）中，所述阳性样本和所述阴性样本的数目比例为 1:0.1~5。

35 18、根据权利要求 1 所述的方法，其特征在于，在步骤（4）中，所述阳性样本和所述

阴性样本的数目比例为 1:0.25~4。

19、根据权利要求 1 所述的方法，其特征在于，所述阳性样本和所述阴性样本针对所述待测染色体以外的其他染色体均不存在非整倍性。

20、根据权利要求 1 所述的方法，其特征在于，在步骤（4）中，采用所述第一特征和
5 所述第二特征确定所述孕妇样本和所述对照样本的二维特征向量，基于由所述二维特征向量确定的样本间距离，将所述孕妇样本在所述阳性对照样本和所述阴性对照样本之间进行归类，以便确定所述胎儿针对所述待测染色体是否存在非整倍性。

21、根据权利要求 20 所述的方法，其特征在于，所述距离为欧几里得距离、曼哈顿距离或切比雪夫距离。

10 22、根据权利要求 20 所述的方法，其特征在于，在步骤（4）中，进一步包括：

（4-1）分别计算所述孕妇样本与所述对照样本之间的距离；

（4-2）将所得到的所述距离进行排序，所述排序基于由小到大的顺序；

（4-3）基于所述排序，从小到大选择预定数量的对照样本；

（4-4）分别确定所述预定数量的所述对照样本中阳性样本和阴性样本的数目；

15 （4-5）基于多数决策法，确定将所述孕妇样本的归类结果。

23、根据权利要求 22 所述的方法，其特征在于，所述预定数量为不超过 20。

24、根据权利要求 22 所述的方法，其特征在于，所述预定数量为 3~10。

25、根据权利要求 22 所述的方法，其特征在于，在步骤（4-2）中，在进行所述排序之前，预先对所述待测样本与预定所述对照样本之间的距离进行加权处理。

20 26、一种确定胎儿是否存在染色体非整倍性的装置，其特征在于，包括：

数据获取模块，用于获取来自孕妇样本的核酸测序数据，所述孕妇样本含有胎儿游离核酸，所述核酸测序数据由多个测序读段构成；

胎儿浓度-反估浓度确定模块，用于基于所述核酸测序数据确定所述孕妇样本的胎儿浓度以及预定染色体的反估浓度，所述反估浓度是基于所述预定染色体的测序读段数目和第一比较染色体的测序读段数目的差异确定的，所述预定染色体包括待测染色体和第二比较
25 染色体，所述第一比较染色体包括至少一个不同于所述预定染色体的常染色体；

特征确定模块，基于所述待测染色体的反估浓度与所述第二比较染色体的反估浓度的差异确定第一特征，用于基于所述待测染色体的反估浓度与所述胎儿浓度的差异确定第二特征；和

30 非整倍性确定模块，用于基于所述第一特征和第二特征，利用对照样本的相应数据，确定所述孕妇的胎儿针对所述待测染色体是否存在非整倍性，其中，所述对照样本包括阳性样本和阴性样本，所述阳性样本针对所述待测染色体具有非整倍性，所述阴性样本针对所述待测染色体不具有非整倍性。

27、根据权利要求 26 所述的装置，其特征在于，所述胎儿浓度-反估浓度确定模块包括：

35 比对单元，用于将来自所述孕妇样本的所述核酸测序数据与参照序列比对，以便确定落入预定窗口的测序读段的数目；和

胎儿浓度计算单元，用于基于所述落入预定窗口的测序读段的数目，确定所述孕妇样本的胎儿浓度。

28、根据权利要求 26 所述的装置，其特征在于，所述胎儿浓度-反估浓度确定模块包括：反估浓度计算单元，用于按照下列公式确定所述反估浓度：

$$F_j = 2 * |R_j - R_r| / (R_r)$$

其中

j 表示需要确定所述反估浓度的染色体的编号，

F_j 表示第 j 号染色体的反估浓度，

R_r 表示多条常染色体的平均测序读段数目，和

10 R_j 表示第 j 号染色体的测序读段数目。

29、根据权利要求 26 所述的装置，其特征在于，所述胎儿浓度-反估浓度确定模块包括：第二比较染色体确定单元用于将多条常染色体的所述反估浓度按照由小至大的优先顺序，选择目标排序的常染色体作为所述第二比较染色体。

30、据权利要求 26 所述的装置，其特征在于，所述特征确定模块包括：

15 第一特征确定单元，用于通过下列公式确定所述第一特征：

$$X_1 = F_i - F_r$$

其中

X_1 表示第一特征，

i 表示所述待测染色体的编号，

20 F_i 表示所述待测染色体的所述反估浓度，

F_r 表示所述第二比较染色体的反估浓度平均值。

31、据权利要求 26 所述的装置，其特征在于，所述特征确定模块包括：

第二特征确定单元，用于通过下列公式确定所述第二特征：

$$X_2 = \frac{F_i}{F_a}$$

25 其中

X_2 表示第二特征，

i 表示所述待测染色体的编号，

F_i 表示所述待测染色体的所述反估浓度，

F_a 表示所述胎儿浓度。

30 32、据权利要求 26 所述的装置，其特征在于，所述特征确定模块包括：

标准化处理单元，用于对所述第一特征和所述第二特征进行标准化处理，以便所述第一特征和所述第二特征的绝对值分别独立地处于 0~1 之间。

33、据权利要求 26 所述的装置，其特征在于，所述非整倍性确定模块用于采用所述第一特征和所述第二特征确定所述孕妇样本和所述对照样本的二维特征向量，基于由所述二
35 维特征向量确定的样本间距离，将所述孕妇样本在所述阳性对照样本和所述阴性对照样本

之间进行归类，以便确定所述胎儿针对所述待测染色体是否存在非整倍性。

34、据权利要求 33 所述的装置，其特征在于，所述距离为欧几里得距离、曼哈顿距离或切比雪夫距离。

35、据权利要求 26 所述的装置，其特征在于，所述非整倍性确定模块用于采用 k-近邻模型确定将所述孕妇样本的归类结果。

36、根据权利要求 35 所述的装置，其特征在于，所述 k-近邻模型采用的 K 值为不超过 20。

37、根据权利要求 35 所述的装置，其特征在于，所述 k-近邻模型采用的 K 值为 3~10。

38、根据权利要求 35 所述的装置，其特征在于，所述 k-近邻模型中，对所述样本间距离进行加权处理。

39、一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现权利要求 1-25 中任一项所述方法的步骤。

40、一种电子设备，其特征在于，包括：

权利要求 39 中所述的计算机可读存储介质；以及
一个或者多个处理器，用于执行所述计算机可读存储介质中的程序。

41、一种构建机器学习分类模型的方法，其特征在于，包括：

(a) 针对多个孕妇样本的每一个分别进行：

获取来自所述孕妇样本的核酸测序数据，所述孕妇样本含有胎儿游离核酸，所述核酸测序数据由多个测序读段构成，所述孕妇样本包括至少一个阳性样本和至少一个阴性样本，所述阳性样本针对待测染色体具有非整倍性，所述阴性样本针对所述待测染色体不具有非整倍性；

基于所述核酸测序数据确定所述孕妇样本的胎儿浓度以及预定染色体的反估浓度，所述反估浓度是基于所述预定染色体的测序读段数目和第一比较染色体的测序读段数目的差异确定的，所述预定染色体包括待测染色体和第二比较染色体，所述第一比较染色体包括至少一个不同于所述预定染色体的常染色体；和

基于所述待测染色体的反估浓度与所述第二比较染色体的反估浓度的差异确定第一特征，基于所述待测染色体的反估浓度与所述胎儿浓度的差异确定第二特征，

(b) 将所述多个孕妇样本作为样本，利用所述样本的第一特征和第二特征，进行机器学习训练，以便构建用于确定胎儿是否具有非整倍性的机器学习分类模型。

42、根据权利要求 41 所述的方法，其特征在于，所述机器学习分类模型为 KNN 模型。

43、根据权利要求 42 所述的方法，其特征在于，所述 KNN 模型采用欧几里得距离。

44、一种构建机器学习分类模型的装置，其特征在于，包括：

特征获取模块，用于针对多个孕妇样本的每一个分别进行：

获取来自所述孕妇样本的核酸测序数据，所述孕妇样本含有胎儿游离核酸，所述核酸测序数据由多个测序读段构成，所述孕妇样本包括至少一个阳性样本和至少一个阴性样

本，所述阳性样本针对待测染色体具有非整倍性，所述阴性样本针对所述待测染色体不具有非整倍性；

- 5 基于所述核酸测序数据确定所述孕妇样本的胎儿浓度以及预定染色体的反估浓度，所述反估浓度是基于所述预定染色体的测序读段数目和第一比较染色体的测序读段数目的差异确定的，所述预定染色体包括待测染色体和第二比较染色体，所述第一比较染色体包括至少一个不同于所述预定染色体的常染色体；和

基于所述待测染色体的反估浓度与所述胎儿浓度的差异确定第二特征，基于所述待测染色体的反估浓度与所述第二比较染色体的反估浓度的差异确定第一特征，

- 10 训练模块，用于将所述多个孕妇样本作为样本，进行机器学习训练，以便构建用于确定胎儿是否具有非整倍性的机器学习分类模型。

45、根据权利要求 44 所述的装置，其特征在于，所述机器学习分类模型为 KNN 模型。

46、一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现权利要求 41~43 任一项所述方法的步骤。