(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 107851465 A
(43)申请公布日 2018.03.27

(21)申请号 201680044286.0

(22)申请日 2016.07.18

(30)优先权数据
62/198,245 2015.07.29 US

(85)PCT国际申请进入国家阶段日
2018.01.29

(86)PCT国际申请的申请数据
PCT/IB2016/054255 2016.07.18

(87)PCT国际申请的公布数据
WO2017/017554 EN 2017.02.02

(71)申请人 皇家飞利浦有限公司
地址 荷兰艾恩德霍芬

(72)发明人 U·拉加万 D·R·埃尔戈特

(74)专利代理机构 永新专利商标代理有限公司 72002
代理人 王英 刘炳胜

(51)Int.Cl.
G16H 70/00(2018.01)

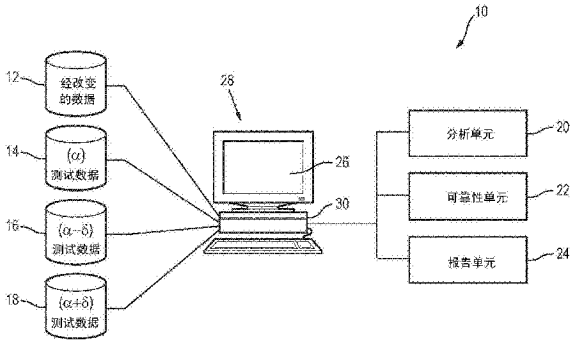
权利要求书2页 说明书6页 附图3页

(54)发明名称

经改变的数据集的数据分析中的可靠性度量

(57)摘要

对经改变的数据进行数据分析包括:使用一个或多个经配置的处理器(30),通过数据分析技术来分析(64)测试数据集(14),其创建一个或多个分析度量,并且所述测试数据集根据置信度分数从经改变的数据集(12)中选择。使用配置的一个或多个处理器,基于所述一个或多个分析度量和根据应用于根据不同置信度分数从经改变的数据集选择的一个或多个可靠性测试数据集(16、18)的数据分析技术创建的相同分析度量的相似性,计算所述一个或多个分析度量的至少一个可靠性度量。



CN 107851465 A

1.一种对经改变数据进行数据分析的方法,包括:

使用一个或多个经配置的处理器(30),利用数据分析技术来分析(64)测试数据集(14),其创建一个或多个分析度量,并且所述测试数据集根据置信度分数从经改变的数据集(12)中选择;

使用所述一个或多个经配置的处理器,基于所述一个或多个分析度量与根据应用于根据不同置信度分数从所述经改变的数据集中选择的一个或多个可靠性测试数据集(16、18)的所述数据分析技术创建的相同分析度量的相似性,来计算(66)所述一个或多个分析度量的至少一个可靠性度量。

2.根据权利要求1所述的方法,其中,所述可靠性度量包括以下中的至少一个:Jaccard指标,学生t测试,Welch的t测试,Kolmogorov-Smirnov测试,或预测性模型准确性度量。

3.根据权利要求1和2中的任一项所述的方法,还包括:

通过改动所述经改变的数据集中的值并且将所述置信度分数与经改动的值相关联,来针对至少一个数据元素改变所述经改变的数据集内的数据。

4.根据权利要求1-3中的任一项所述的方法,还包括:

通过对来自至少两个来源的记录进行匹配将数据整合到所述经改变的数据集中,并将所述置信度分数与整合的数据相关联。

5.根据权利要求1-4中的任一项所述的方法,其中,所述分析度量包括描述性统计、预测性准确性度量、分类或数据分布中的至少一个。

6.根据权利要求1-5中的任一项所述的方法,其中,计算至少一个可靠性度量包括:

基于对从所述经改变的数据集中选择的具有不同于所述置信度分数的第一置信度分数的第一可靠性测试数据集(16)的所述数据分析,来计算第一可靠性度量;并且

基于对从所述经改变的数据中选择的具有不同于所述置信度分数和所述第一置信度分数的第二置信度分数的第二可靠性测试数据集(18)的所述数据分析,来计算第二可靠性度量。

7.根据权利要求6所述的方法,其中,所述第一置信度分数与所述置信度分数相比具有负差值,所述第二置信度分数与所述置信度分数相比具有正差值。

8.根据权利要求1-7中的任一项所述的方法,其中,所述至少一个可靠性度量包括作为个体可靠性度量的函数的复合度量。

9.根据权利要求1-8中的任一项所述的方法,其中,所述至少一个可靠性度量被进一步归类。

10.根据权利要求1-9中的任一项所述的方法,其中,利用所述数据分析技术来分析所述测试数据集包括:

将所述数据分析技术并行地应用于所述测试数据集和所述一个或多个可靠性测试数据集(16、18)。

11.根据权利要求1-10中的任一项所述的方法,还包括:

将所述可靠性分析输出(68)到显示设备、打印设备或计算机文件中的一个。

12.一种用于对经改变的数据进行数据分析的系统(10),包括:

包括一个或多个经配置的处理器的分析单元(20),所述分析单元利用数据分析技术来分析根据置信度分数从经改变的数据集(12)中选择的测试数据集(14),其创建一个或多个

分析度量，并且根据应用于根据不同置信度分数从所述经改变的数据集中选择的一个或多个可靠性测试数据的所述数据分析技术创建相同分析度量；

包括所述一个或多个经配置的处理器的可靠性单元(22)，所述可靠性单元基于所述一个或多个分析度量与应用于所述一个或多个可靠性度量数据集的所述相同分析度量的相似性，来计算所述一个或多个分析度量的至少一个可靠性度量。

13.根据权利要求12所述的系统，其中，所述可靠性度量包括以下中的至少一个：Jaccard指标，学生t测试，Welch的t测试，Kolmogorov-Smirnov测试，或预测性模型准确性测量。

14.根据权利要求12和13中的任一项所述的系统，其中，所述置信度分数根据变动的数据值与所述经改变的数据集相关联。

15.根据权利要求12-13中的任一项所述的系统，其中，根据通过对来自至少两个源的记录进行匹配而被整合到所述经改变的数据集中的数据，所述置信度分数与所述经改变的数据相关联。

16.根据权利要求12-14中的任一项所述的系统，其中，所述分析度量包括描述性统计、预测性准确性度量、分类或数据分布中的至少一个。

17.根据权利要求12-15中的任一项所述的系统，其中，所述可靠性单元基于对从所述经改变的数据集中选择的具有不同于所述置信度分数的第一置信度分数的第一可靠性测试数据集(16)的所述数据分析来计算第一可靠性度量，并且基于对从所述经改变的数据中选择的具有不同于所述置信度分数和所述第一置信度分数的第二置信度分数的第二可靠性测试数据集(18)的所述数据分析来计算第二可靠性度量。

18.根据权利要求12-17中的任一项所述的系统，其中，所述可靠性单元对所述至少一个可靠性度量进行归类。

19.根据权利要求12-18中的任一项所述的系统，其中，所述分析将所述数据分析技术并行地应用于所述测试数据集和所述一个或多个可靠性测试数据集。

20.一种对经改变的数据进行数据分析的方法，包括：

从经改变的数据集中选择具有大于阈值量的第一置信度分数的测试数据集、与所述第一置信度分数相比具有负差值的第二置信度分数的第一可靠性测试数据集、以及与所述第一置信度分数相比具有正差值的第三置信度分数的第二可靠性测试集；

利用使用一个或多个处理器应用的数据分析技术来分析所述测试数据集、所述第一可靠性测试数据集和所述第二可靠性测试数据集，其创建一组分析度量，针对每个被分析的数据集至少一个分析度量；

基于来自经分析的测试数据集的所述至少一个分析度量和来自经分析的第一可靠性测试数据集的所述至少一个分析度量，来计算所述至少一个分析度量的第一可靠性度量值；并且基于来自经分析的测试数据集的所述至少一个分析度量和来自经分析的第二可靠性测试数据集的所述至少一个分析度量，来计算所述至少一个分析度量的第二可靠性度量。

经改变的数据集的数据分析中的可靠性度量

技术领域

[0001] 以下总体涉及数据分析和数据挖掘，具体应用于通过健康护理数据的数据清理和数据整合所改变的数据集的数据分析。

背景技术

[0002] 数据挖掘已经在具有从多种来源累积的数据的大数据集上进行。数据挖掘可以包括数据收集、数据结构化、数据清理(例如，消除不一致)、纠正错误、整合或编译来自不同来源的数据，并分析数据以获得新的信息。来自健康护理提供者的数据可以提供关于患者风险、健康护理处置或趋势的信息。诸如聚类分析、方差分析和其他统计技术的数据分析通常接受数据值是准确的，并且通过识别和去除离群值来专注于归类/分类/预测。

[0003] 当数据被修改以准备分析时，数据的改变会给数据增加不确定性，从而可能带来对不确定数据的分析。例如，药物名称可能拼写错误，使用商品名称，使用缩写等。一种方法是在数据清理期间标记任何改变的数据。基于通过数据清理修改的识别组中记录的百分比来判断后续分析的可靠性，例如，根据聚类分析在识别出的聚类中的修改数据的高百分比指示该聚类是可疑的。然而，使用标记不会区分数据改变类型，其中一些是显而易见的，例如小的拼写错误，而一些不太明显，如缩写或替代名称。清理数据的过程可以在经清理的数据中引入新的模式，这被认为是虚假的，例如，指示清理过程，并不反映原始数据或底层数据模式。

[0004] 对随后分析数据引入不确定性的另一领域是对来自不同来源的数据的整合。健康护理提供者按规章提供去识别的患者数据，即从数据中去除患者身份。数据来源可以包括来自健康护理提供者内的不同区域，例如患者护理记录、帐单、入院、药房、放射科等。源可以在不同的健康护理提供者之间，例如不同的站点，不同的医院，不同的门诊等。由于数据是从不同的来源整合来识别模式，所以匹配算法会增加不确定性，并将其传递给后续的分析。例如，去识别的患者诊断可以与去识别的药房记录整合。根据诊断处方的药物的分析可以包括根据患者诊断如何匹配到药房记录的错误，例如虚假的，而不是基于诊断如何对患者进行药物处方，例如不是虚假的。然而，数据分析技术不包括数据整合的可靠性度量，通常只有用于所应用数据分析技术的置信度分数或准确性测量，例如回归分析/方差分析中的R^2值。

发明内容

[0005] 本文描述的各方面解决了上述问题和其他问题。

[0006] 以下描述了一种确定对经改变的数据的分析的可靠性度量的方法和系统。所述经改变的数据包括与数据相关联的置信度分数。置信度分数可以与通过数据清理和/或通过数据整合而整合的记录实例而改变的数据元素的特定实例相关联。

[0007] 在一个方面，一种分析技术使用配置的一个或多个处理器，所述配置的一个或多个处理器创建一个或多个分析度量，并且测试数据集根据置信度分数从经改变的数据集中

选择。使用配置的一个或多个处理器，基于所述一个或多个分析度量与根据应用于根据不同置信度分数从经改变的数据集选择的一个或多个可靠性测试数据集的数据分析技术创建的相同分析度量的相似性，计算所述一个或多个分析度量的至少一个可靠性度量。

[0008] 在另一方面，一种用于对经改变的数据进行数据分析的系统包括分析单元和可靠性单元。所述分析单元包括一个或多个配置的处理器，所述分析单元利用数据分析技术来分析根据置信度分数从经改变的数据集选择的测试数据集，其创建一个或多个分析度量，并且根据被应用到根据不同置信度分数从经改变的数据集选择的一个或多个可靠性测试数据集的数据分析技术创建相同分析度量。所述可靠性单元包括一个或多个经配置的处理器，所述可靠性单元基于所述一个或多个分析度量和应用于所述一个或多个可靠性度量数据集的相同分析度量的相似性，来计算所述一个或多个分析度量的至少一个可靠性度量。

[0009] 在另一方面，一种对经改变的数据进行数据分析的方法，包括：从经改变的数据集中选择具有大于阈值量的第一置信度分数的测试数据集，与所述第一置信度分数相比具有负差值的第二置信度分数的第一可靠性测试数据集，以及与所述第一置信度分数相比具有正差值的第三置信度分数的第二可靠性测试集。通过使用一个或多个处理器应用的数据分析技术来分析所述测试数据集、所述第一可靠性测试数据集和所述第二可靠性测试数据集，所述一个或多个处理器创建一组分析度量，至少一个分析度量用于被分析的每个数据集。基于来自经分析的测试数据集的所述至少一个分析度量和来自经分析的第一可靠性测试数据集的所述至少一个分析度量，来计算所述至少一个分析度量的第一可靠性度量值；并且基于来自经分析的测试数据集的所述至少一个分析度量和来自经分析的第二可靠性测试数据集的所述至少一个分析度量，来计算所述至少一个分析度量的第二可靠性度量。

附图说明

[0010] 本发明可以采取各种部件和部件布置以及各种步骤和步骤安排的形式。附图仅用于说明优选实施例的目的，而不应被解释为限制本发明。

[0011] 图1示意性地图示了在经改变的数据集系统的数据分析中的可靠性度量的实施例。

[0012] 图2图示了具有数据分析的可靠性度量的示例性报告。

[0013] 图3是在经改变的数据集的数据分析中的可靠性度量的实施例的流程图。

具体实施方式

[0014] 首先参考图1，示意性地图示了经改变的数据集系统10的数据分析中的可靠性度量的实施例。系统10包括经改变的数据集12或对经改变的数据集12的电子访问，从其可以导出测试数据集14和一个或多个可靠性测试数据集16、18。经改变的数据集12包括一个或多个数据元素和/或包括相关联的置信度分数的记录。相关联的置信度分数可以通过数据清理和/或数据整合来关联。置信度分数可以表示为连续范围的值，例如，0.1-100.0、0.01-1.00、1-100等。

[0015] 例如，处方药物名称Propofal、Diprivan、Fospropofol和Propofol的出现被确定为与数据集中Propofol相同的药物名称。药物的名称是处方药物的数据元素或属性。通过数据清理，将药物名称的不同出现改变为Propofol，并且与以下置信度相关联：(Propofal

到Propofol)98%,(Diprivan到Propofol)99%,(Fospropofol到Propofol)25%,以及100%(未改变)。在经改变的数据集中的数据元素“药物名称”中出现“Propofol”包括指示名称改变代表真实信息的置信度的相关联置信度分数。相关联的置信度分数可以在记录水平上被存储,例如,附加到实例或发生,或单独存储,例如链接或相关表。记录包括一组相关的数据元素,例如,患者的属性。示例性技术在2014年12月9日提交的、序列号为62/089336的、标题为“System and Method for Uniformly Correlating Unstructured Entry Features to Associated Therapy Features”的专利申请中更全面地描述,在此通过引用将其全部内容并入。

[0016] 通过数据整合相关联的置信度分数与纪录水平相关联。例如,包括以下数据元素和值的第一数据源:年龄=63,性别=f,种族=亚洲人,诊断=AMI,HR=30,匹配于包括以下数据元素和值的第二数据源:年龄=64,性别=f,种族=亚洲人,诊断=AMI,总费用=$12,340,结果=30天再入院。该匹配与73%的置信度分数相关联,该置信度分数指示该匹配有效的置信度,例如,该匹配是同一患者。通过具有以上值的年龄、性别、种族、诊断、HR、总费用和结果的组合数据元素所识别出的患者的发生与73%的置信度分数相关联。其他匹配或发生可以是不同的值。示例性技术在2015年2月27日提交的、序列号为62/121,608、标题为“Efficient Integration of De-Identified Records”的专利申请中更全面地描述,在此将其全部内容并入。

[0017] 测试数据集14包括至少一个数据元素,其具有基于置信度度量中的一个从经改变的数据集12中选择的发生。例如,选择与“药品名称”相关联的置信度分数大于75%的发生。测试数据集14可以包括来自经改变的数据集的数据元素的子集。例如,测试数据集包括年龄、性别、诊断、HR,并且整合置信度分数的结果是80%或更大,即a≥80%,其中“a”是记录发生的置信度分数,不包括“总费用”数据元素。在另一例子中,测试数据集包括年龄、性别、药物名称和诊断,其中药物名称的置信度度量为75%或更多,例如,≥75%。可靠性测试数据集16、18包括基于数据分析并具有变化的置信度水平(α±δ)的相同数据元素。可以使用本领域已知的数据操纵技术从经改变的数据集12提取或创建测试数据集14和可靠性测试数据集16、18。在一个实施例中,系统10基于所选的数据元素和用户可修改的默认置信水平来生成测试数据集14,并且生成具有置信水平中用户可修改的默认差异的可靠性测试数据集16、18。在一个实施例中,数据分析单元20执行数据集创建或提取。

[0018] 数据分析单元20或用户使用已知的数据分析技术来应用数据分析,所述数据分析技术例如是描述和/或汇总统计、关联分析、聚类分析、分类、预测分析等。数据分析技术应用于测试数据集14。例如,数据分析单元将聚类分析应用于所选的置信度分数大于80%的年龄、体重(kg)、心率(以每分钟心跳为单元的HR)以及肌酐的测试数据集,例如数据整合相关联置信度分数>a。将相同的数据分析应用于每个可靠性测试数据集16、18。在一个实施例中,利用测试数据集12分析来自动执行可靠性测试数据集16、18的生成和分析。在另一实施例中,基于用户提示或用户输入,在分析测试数据集14之后执行可靠性测试数据集16、18的生成和分析,以执行可靠性测试。

[0019] 可靠性单元22基于测试数据集12和可靠性测试数据集16、17的数据分析来计算可靠性度量,例如用于聚类分析的Jaccard指标,用于描述统计的t测试,用于预测分析的R^2值,等等。例如,假设聚类C_1、C_2和C_3是在测试数据集12上应用k均值聚类算法的结果,聚类

C_{11}、C_{12}、C_{13}是在第一可靠性测试数据集16(X1)上应用k均值聚类算法的结果，并且假设聚类C_{21}、C_{22}、C_{23}是对第二可靠性测试数据集18(X2)应用k均值聚类算法的结果。计算Jaccard指标，用于比较$\{C_{11},C_{12},C_{13}\}$与限于X1的记录的原始聚类$\{C_1,C_2,C_3\}/_{X1}$。如果r代表两个集合中同一聚类中的数据点对，s代表X中同一聚类中但在X_1中的不同聚类中的数据点对，并且及t代表X_1中同一聚类但在X中的不同聚类中的数据点对，则Jaccard指标被定义为(r/(r+s+t))。如果指数是1，则两组聚类是相同的，当指数是0时，它们是完全不相似的。接近1的值可以指示两个解决方案之间的强相似性。针对第二测试数据集18(X2)计算Jaccard指标。可靠性度量(例如Jaccard指标)可以包括一系列的值，如0-100，或者可靠性度量可以根据计算出的测量进行归类。

[0020] 在另一例子中，例如描述性统计，使用学生t测试或Welch的t测试，在测试数据集12和可靠性数据集16、18之间比较均值和/或标准差。例如，t测试计算两个均值是相同真均值的可能性。如果空假设是两个均值是不同均值，并且针对测试数据集和第一可靠性测试数据集的均值的t测试比较不被拒绝，并且针对测试数据集和第二可靠性测试数据集的均值的t测试比较也不被拒绝，则结果是将复合可靠性度量归类为虚假的。如果空假设针对测试数据集和第一可靠性测试数据集的t测试不被拒绝，并且针对测试数据集和第二可靠性测试数据集的t测试被拒绝，则结果是归类为可能是虚假的。如果空假设针对两个比较都被拒绝，则结果被归类为可靠的。

[0021] 可以使用Kolmogorov-Smirnov测试来比较数据集的分布，例如，每个数据集的分布表示相同分布的可能性。预测模型可以使用精确性度量(例如R^2值)进行比较。例如，通过相同的预测子或自变量，R^2的比较提供了模型拟合的相似性的指示。

[0022] 可靠性单元22可以将可靠性度量组合或归类为复合度量。在一个实施例中，可靠性度量可以被归类或解释为归类度量，例如“可靠”、“可能是虚假的”、“绝对是虚假的”。例如，0.0-1.0范围的Jaccard指标可以被归类为0.0-0.39，虚假的；0.4-0.69，可能是虚假的；以及0.7-1.0，可靠的。例如，使用预测度量，相对差异：$(R^2(X)-R^2(X_1))/(R^2(X))$的变化大于50％可以被归类为虚假的，在5％和50％之间可能是虚假的，并且低于5％是可靠的。归类范围和置信度分数可以根据用户偏好、系统默认和/或项目偏好等来设置。

[0023] 报告单元24显示数据分析的结果和可靠性度量。例如，显示可以被打印或显示在显示设备26上，例如计算机设备28的显示器。显示器可以包括原始可靠性度量、复合度量和/或归类度量。

[0024] 分析单元20、可靠性单元22和报告单元24包括至少一个处理器30(例如，微处理器、中央处理单元、数字处理器等)，其被配置为执行存储于计算机可读存储介质中的至少一条计算机可读指令，所述存储介质不包括瞬态介质并且包括物理存储器和/或其他非瞬态介质。处理器30还可以执行由载波、信号或其他瞬态介质承载的一条或多条计算机可读指令。处理器30可以包括本地存储器和/或分布式存储器。处理器30可以包括用于有线和/或无线通信的硬件/软件。处理器30可以包括计算设备28，例如台式计算机、服务器、膝上型计算机、移动设备、分布式设备、其组合等。

[0025] 参考图2，图示了具有数据分析的可靠性度量的示例性报告。示例性报告包括数据分析40的报告，其是对从经改变的数据集12选择的置信水平(>a)的测试数据集14的聚类分析。聚类分析指示具有数据元素或属性的三个识别出的聚类：以年为单位的年龄，以千克

(kg)为单位的体重,以每分钟搏动次数(bpm)为单位的心率,以及以毫克/分升(mg/dl)为单位的肌酐。第一聚类包括分别针对年龄、体重、心率和肌酐的62、92、70和1.1的值。第二聚类包括分别为71、94、65和1.5的值,第三聚类包括分别为77、71、50和3.9的值。

[0026] 示例性报告包括测试数据集14与第一可靠性测试数据集16的相似性的可靠性度量44,其被呈现为归类为中等或可能是虚假的。第二可靠性度量46指示测试数据集14与第二可靠性测试数据集18之间的相似性,其被归类为差的或者绝对是虚假的。示出了复合度量48,其绝对是虚假的。图例50表示可靠的、可能是虚假的以及绝对是虚假的不同类别。

[0027] 因此,根据具有可靠性度量44、46、48的示例性报告,用户可以合理地推断形成的三个聚类可能归因于作为数据清理和/或数据整合的结果而引入的模式,而不是表示真实的底层的数据的模式。

[0028] 参考图3,用流程图图示了经改变的数据集12的数据分析中的可靠性度量的实施例。在60处,接收经改变的数据集12,其包括针对至少一个数据元素或一组记录的置信度分数。经改变的数据集12可以通过参考例如计算机存储器和/或存储设备中的位置的识别,或通过电子传输(例如,通过网络连接从一个存储位置传送到另一个)来接收。在一个实施例中,接收可以包括清理数据并将置信度分数分配给经清理/经更改的数据。在一个实施例中,接收可以包括整合两个或更多个数据源并将置信度分数分配给整合数据,例如,匹配或组合的记录。在另一实施例中,接收可以包括数据清理和数据整合的组合。

[0029] 在62处,通过从置信度分数高于预定阈值的经改变的数据集12中选择数据而生成测试数据集14。例如,选择包括药物名称的一组数据元素,其中,与药物名称相关联的置信度分数大于70%,例如,$\alpha>70\%$。在另一例子中,从经改变的数据集中选择一组数据元素,其中,与整合记录相关联的置信度分数大于75%。

[0030] 在64处,具有高于预定量的置信度分数(α)的测试数据集14由分析单元20使用数据分析技术进行分析。数据分析输出测试数据集14的至少一个分析度量,例如聚类、均值、标准偏差、R^2值、类别等。

[0031] 在66处计算可靠性度量,其评估测试数据分析的可靠性。根据对选择具有与测试数据集12相同的数据元素的第一可靠性数据集16的相同分析的输出分析度量和与预定分数($\alpha-\delta$)具有负差值的置信度分数,来计算可靠性度量,并且输出第二可靠性数据集18的相同分析的置信度分数与预定分数($\alpha-\delta$)成正差的分析度量。可靠性度量包括输出分析度量的相似性的原始度量,例如Jaccard指标、T测试等。可靠性度量可以被归类和/或组合成复合度量。在一个实施例中,响应于来自测试数据集14的分析的显著输出分析度量来计算可靠性数据集16、18的分析度量和可靠性度量。在另一实施例中,分析度量计算与测试数据集14分析并行进行,并且在分析度量的输出之后计算可靠性度量。

[0032] 在68处,报告可靠性度量。可靠性度量可作为原始度量、归类的原始度量、复合度量或归类的复合度量进行报告。报告可以与显示设备上的测试数据集14的输出分析度量一起呈现,或者合并在电子或打印文件中以供后续查看。

[0033] 以上可以通过编码或嵌入在计算机可读存储介质上的计算机可读指令来实现,所述计算机可读指令在由(一个或多个)计算机处理器执行时使所述(一个或多个)处理器执行所描述的动作。另外或替代地,计算机可读指令中的至少一个由信号、载波或其他瞬态介质承载。

[0034] 已经参照优选实施例描述了本发明。他人在阅读和理解了前面的详细描述后，可能会想到修改和变化。目的是，将本发明构造为包括所有这样的修改和变化，只要它们落入所附权利要求或其等价方案的范围内。

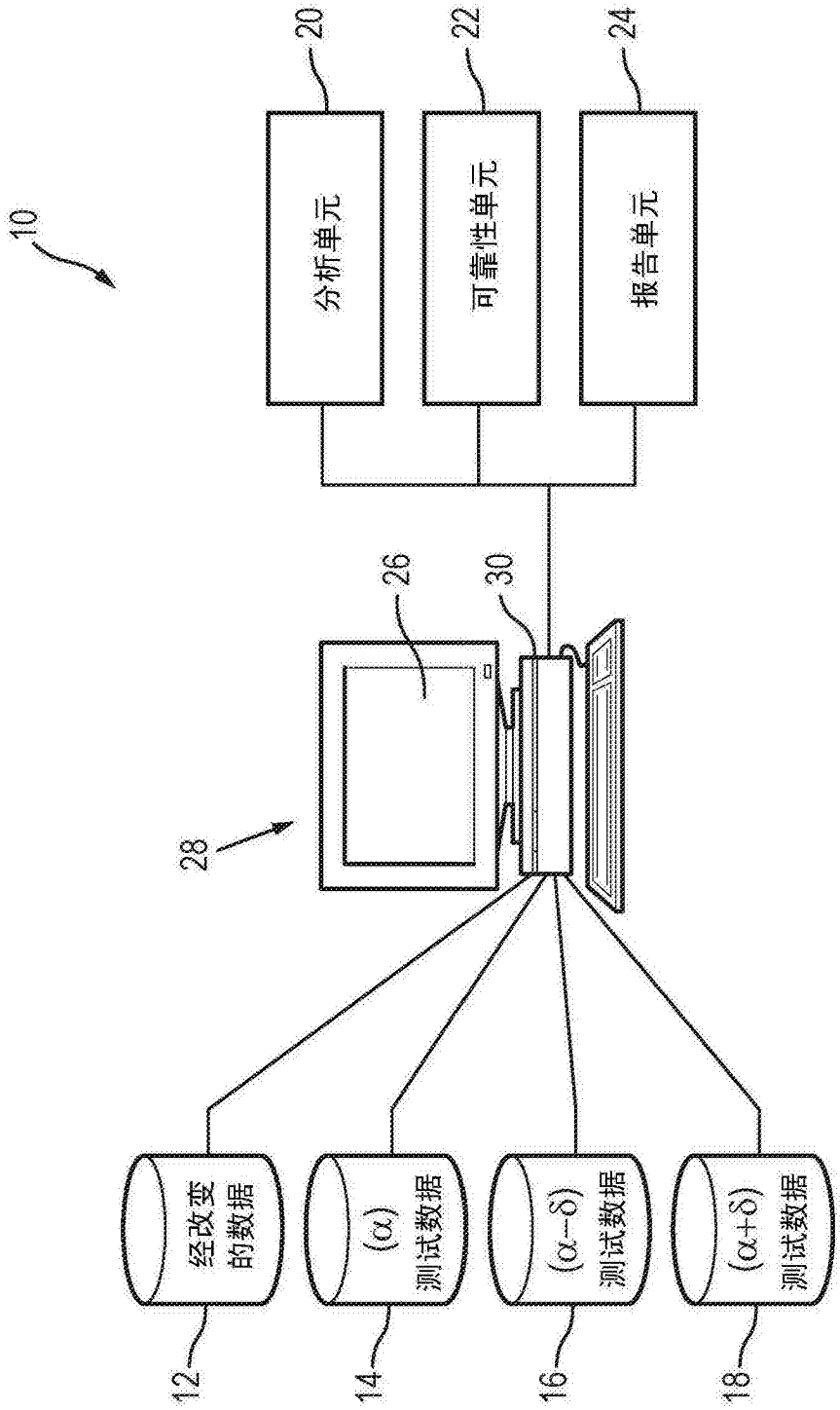

10
20
分析单元
22
可靠性单元
24
报告单元
26
30
28
经改变的数据
12
(α)测试数据
14
(α−δ)测试数据
16
(α+δ)测试数据
18

图1

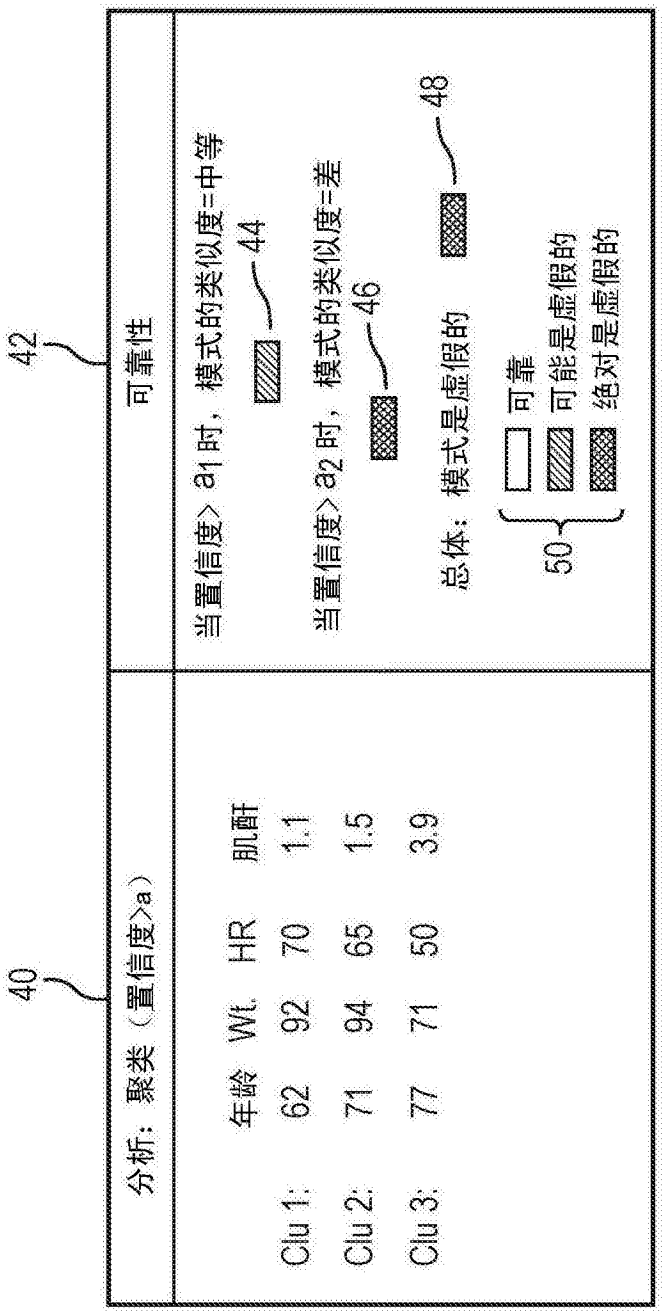

40
分析：聚类（置信度>a）
年龄 Wt. HR 肌酐
Clu 1: 62 92 70 1.1
Clu 2: 71 94 65 1.5
Clu 3: 77 71 50 3.9
42
可靠性
当置信度> a1 时，模式的类似度=中等
44
当置信度> a2 时，模式的类似度=差
46
总体：模式是虚假的
48
50
可靠
可能是虚假的
绝对是虚假的

图2

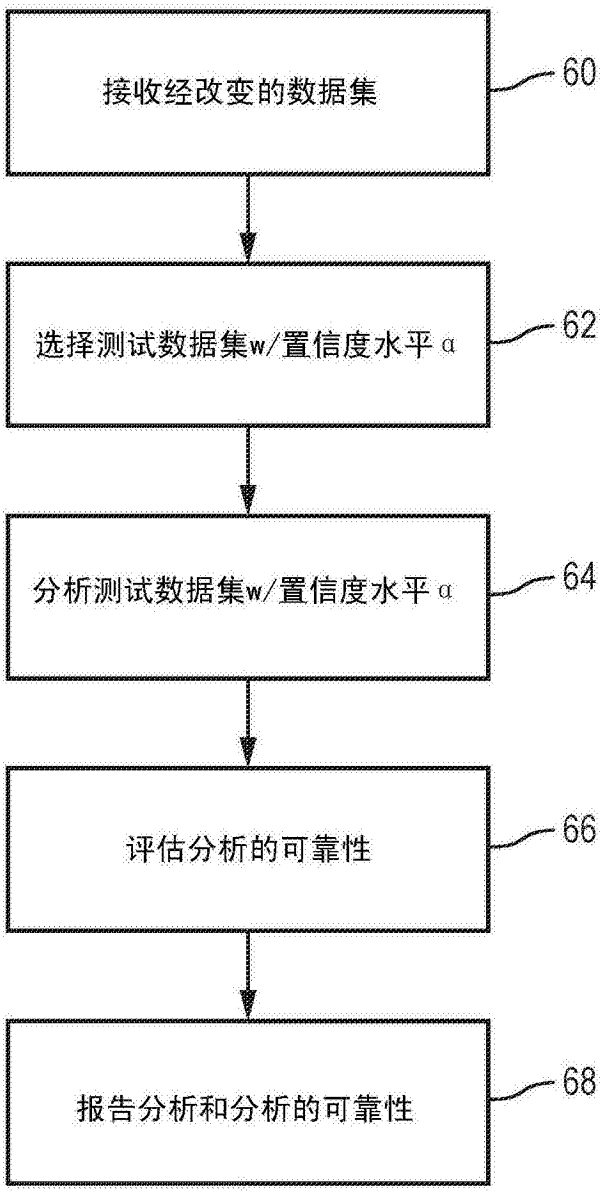

接收经改变的数据集
60
选择测试数据集w/置信度水平α
62
分析测试数据集w/置信度水平α
64
评估分析的可靠性
66
报告分析和分析的可靠性
68

图3