(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115362452 A
(43)申请公布日 2022.11.18

(21)申请号 202180023331.5

(22)申请日 2021.02.24

(30)优先权数据
16/830,905 2020.03.26 US

(85)PCT国际申请进入国家阶段日
2022.09.22

(86)PCT国际申请的申请数据
PCT/IB2021/051532 2021.02.24

(87)PCT国际申请的公布数据
WO2021/191703 EN 2021.09.30

(71)申请人 国际商业机器公司
地址 美国纽约

(72)发明人 R·比加尔 L·克米埃洛斯基
P·斯洛维库斯基 W·索巴拉

(74)专利代理机构 中国贸促会专利商标事务所有限公司 11038
专利代理师 吴信刚

(51)Int.Cl.
G06N 3/08(2006.01)

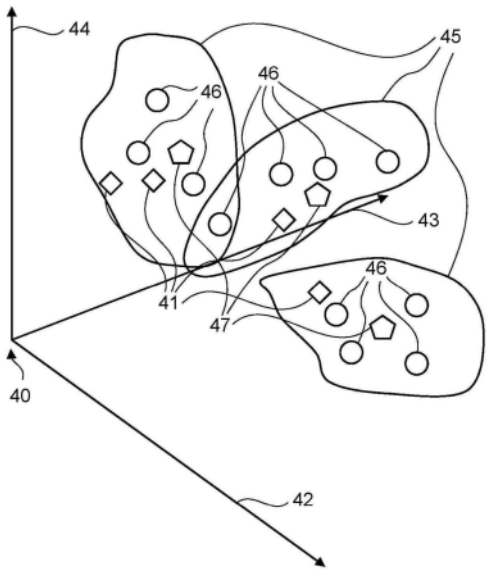
权利要求书3页 说明书16页 附图4页

(54)发明名称

用于选择用于更新人工智能模块的数据集的方法

(57)摘要

一种用于从给定数据集中选择数据集以更新人工智能模块(AI模块)的计算机实现的方法。所述给定数据集(14)中的每个包括输入数据集(11)和对应的输出数据集(12)。该计算机实现的方法包括:获得用于定义所述给定数据集(14)的不同聚类(45)的参数值(301),确定每个给定数据集(14)的度量,每个给定数据集(14)的度量取决于相应给定数据集(14)的针对聚类(45)之一的隶属度以及相应给定数据集(14)到聚类(45)中的同一个聚类的质心(47)的距离(302),以及基于给定数据集(14)的所述度量的比较,从给定数据集(14)中选择给定数据集(14)中的至少一个以用于更新AI模块(1)(303)。



CN 115362452 A

1.一种用于从给定数据集中选择数据集以更新人工智能模块(AI模块)的计算机实现的方法,所述给定数据集中的每个包括输入数据集和对应的输出数据集,所述方法包括:

获得用于定义所述给定数据集的不同聚类的参数值;

确定每个给定数据集的度量,每个给定数据集的度量取决于相应给定数据集的针对所述聚类中的一个聚类的隶属度以及相应给定数据集到所述聚类中的同一个聚类的质心的距离;以及

基于所述给定数据集的所述度量的比较,从所述给定数据集中选择所述给定数据集中的至少一个以用于更新所述AI模块。

2.根据权利要求1所述的计算机实现的方法,还包括:

确定每个聚类的度量,每个聚类的度量取决于相应聚类的质心到其他聚类质心的距离;

基于所述聚类的所述度量从所述聚类中选择所述聚类中的至少一个;以及

确定每个给定数据集的所述度量,每个给定数据集的所述度量取决于相应给定数据集的针对所选聚类的隶属度以及相应给定数据集到所选聚类的质心的距离。

3.根据权利要求1所述的计算机实现的方法,还包括至少部分地基于以下操作来确定每个给定数据集的所述度量:

确定每个给定数据集的度量集合,所述相应给定数据集的度量集合中的每个度量对应于所述聚类的子集中的一个聚类,所述相应给定数据集的度量集合中的每个度量取决于所述相应给定数据集的针对对应聚类的隶属度以及所述相应给定数据集到所述对应聚类的质心的距离;以及

基于所述给定数据集的所述度量集合的比较,从所述给定数据集中选择所述给定数据集中的至少一个以用于更新所述AI模块。

4.根据权利要求1所述的计算机实现的方法,还包括:

根据训练数据集来生成用于定义所述聚类的所述参数值,所述AI模块是使用所述训练数据集生成的。

5.根据权利要求1所述的计算机实现的方法,还包括:

根据所述给定数据集生成用于定义所述聚类的所述参数值。

6.根据权利要求1所述的计算机实现的方法,还包括:

根据测试数据集来生成用于定义所述聚类的所述参数值,所述AI模块使用所述测试数据集被测试。

7.根据权利要求1所述的计算机实现的方法,还包括:

根据所述给定数据集的经批准或校正的数据集来生成用于定义所述聚类的所述参数值。

8.根据权利要求1所述的计算机实现的方法,还包括:

根据所述给定数据集的手动批准或手动校正的数据集来生成用于定义所述聚类的所述参数值。

9.根据权利要求1所述的计算机实现的方法,还包括:

获得用于定义执行模糊C均值聚类算法的所述聚类的所述参数值。

10.根据权利要求2所述的计算机实现的方法,还包括:

基于所述给定数据集到相应聚类的质心的平均距离来确定每个聚类的所述度量。

11.根据权利要求2所述的计算机实现的方法,还包括:

基于所述给定数据集到相应聚类的质心的最大距离来确定每个聚类的所述度量。

12.根据权利要求2所述的计算机实现的方法,还包括:

基于所述给定数据集的针对相应聚类的平均隶属度来确定每个聚类的所述度量。

13.根据权利要求4所述的计算机实现的方法,还包括:

基于所述训练数据集和所述给定数据集的手动批准或手动校正的数据集到相应聚类的质心的平均距离来确定每个聚类的所述度量。

14.根据权利要求4所述的计算机实现的方法,还包括:

基于所述训练数据集和所述给定数据集的手动批准或手动校正的数据集到相应聚类的质心的最大距离来确定每个聚类的所述度量。

15.根据权利要求4所述的计算机实现的方法,还包括:

基于所述训练数据集和所述给定数据集的手动批准或手动校正的数据集的针对相应聚类的平均隶属度来确定每个聚类的所述度量。

16.根据权利要求4所述的计算机实现的方法,还包括:

基于由相应聚类包括的所述训练数据集的数量与由相应聚类包括的所述给定数据集的手动批准或手动校正的数据集的数量的第一和与所述训练数据集的总数与所述给定数据集的手动批准或手动校正的数据集的总数的第二和的比率,确定每个聚类的所述度量。

17.根据权利要求4所述的计算机实现的方法,还包括:

基于所述训练数据集的输出数据集来获得用于定义所述聚类的所述参数值。

18.根据权利要求1所述的计算机实现的方法,其中所述给定数据集的所述输入数据集中的每个包括标识参数的值,并且所述给定数据集的所述输出数据集中的每个包括性能指标的值。

19.一种用于从给定数据集中选择数据集以更新人工智能模块(AI模块)的计算机程序产品,所述给定数据集中的每个包括输入数据集和对应的输出数据集,所述计算机程序产品包括具有随其实施的计算机可读程序代码的计算机可读存储介质,所述计算机可读程序代码被配置成实现一种方法,该方法包括:

获得用于定义所述给定数据集的不同聚类的参数值;

确定每个给定数据集的度量,每个给定数据集的度量取决于相应给定数据集的针对所述聚类中的一个聚类的隶属度以及相应给定数据集到所述聚类中的同一个聚类的质心的距离;以及

基于所述给定数据集的所述度量的比较,从所述给定数据集中选择所述给定数据集中的至少一个以用于更新所述AI模块。

20.一种用于从给定数据集中选择数据集以更新人工智能模块(AI模块)的计算机系统,所述给定数据集中的每个包括输入数据集和对应的输出数据集,所述计算机系统包括一个或多个计算机处理器、一个或多个计算机可读存储介质、以及存储在所述一个或多个计算机可读存储介质上以供所述一个或多个计算机处理器执行来实现一种方法的程序指令,所述方法包括:

获得用于定义所述给定数据集的不同聚类的参数值;

确定每个给定数据集的度量,每个给定数据集的度量取决于相应给定数据集的针对所述聚类中的一个聚类的隶属度以及相应给定数据集到所述聚类中的同一个聚类的质心的距离;以及

基于所述给定数据集的所述度量的比较,从所述给定数据集中选择所述给定数据集中的至少一个以用于更新所述AI模块。

用于选择用于更新人工智能模块的数据集的方法

背景技术

[0001] 本发明涉及数字计算机系统领域,更具体地,涉及一种用于选择数据集以适应人工智能模块的方法。

[0002] 人工智能(AI)或机器智能是感知其环境并采取使其成功实现目标的机会最大化的动作的任何设备。人工智能通常被理解为模仿人类与人类头脑相关联的“认知”功能的机器或计算机,所述“认知”功能例如为语音识别、学习、推理、规划和问题解决。机器学习(人工智能的子集)允许设备自动地从过去的数据中学习,而不使用明确的指令,而是依赖于模式和推断。机器学习算法基于样本数据(称为“训练数据”)建立数学模型,以便在没有明确编程以执行任务的情况下做出预测或决策。当新的训练数据变得可用时,更新或重新训练机器学习算法。

发明内容

[0003] 在应用经训练的人工智能模块(AI模块)的过程中,可能发生的是,其目的在于改进AI模块。这种改进可通过使用尚未用于训练或验证AI模块的附加数据集来更新、优选地重新训练AI模块来执行。这些附加数据集可通过将应用于AI模块的输入数据集记录到日志文件中并将由AI模块基于输入数据集计算的相应输出数据集记录到日志文件中来收集。

[0004] 本发明的各实施例提供了如独立权利要求的主题所描述的用于从给定数据集中选择用于更新人工智能模块(AI模块)的数据集的计算机实现的方法、计算机程序产品和计算机系统。在从属权利要求中描述了有利的实施例。如果本发明的实施例不是相互排斥的,则它们可以彼此自由地组合。

[0005] 根据一个实施例,本发明包括一种用于从给定数据集中选择数据集以更新人工智能模块(AI模块)的计算机实现的方法,所述给定数据集各自包括输入数据集和对应的输出数据集。该计算机实现的方法包括:获得用于定义给定数据集的不同聚类的参数值,确定每个给定数据集的度量,每个给定数据集的度量取决于相应给定数据集的针对聚类之一的隶属度(level of membership)以及相应给定数据集到所述聚类中的同一个聚类的质心的距离,以及基于给定数据集的度量的比较从给定数据集中选择给定数据集中的至少一个以用于更新AI模块。

[0006] 根据另一实施例,本发明包括一种用于从给定数据集选择数据集以更新人工智能模块(AI模块)的计算机程序产品,所述给定数据集各自包括输入数据集和对应的输出数据集,所述计算机程序产品包括计算机可读存储介质,所述计算机可读存储介质具有随其实施的计算机可读程序代码,所述计算机可读程序代码被配置成实现一种方法,所述方法包括:获得用于定义给定数据集的不同聚类的参数值,确定每个给定数据集的度量,每个给定数据集的度量取决于相应给定数据集的针对聚类之一的隶属度以及相应给定数据集到所述聚类中的同一个聚类的质心的距离,以及基于给定数据集的度量的比较从给定数据集中选择给定数据集中的至少一个以用于更新AI模块。

[0007] 根据另一实施例,本发明包括一种用于从给定数据集选择数据集以更新人工智能

模块(AI模块)的计算机系统,所述给定数据集各自包括输入数据集和对应的输出数据集,所述计算机系统包括一个或多个计算机处理器、一个或多个计算机可读存储介质、以及存储在所述一个或多个计算机可读存储介质上以供所述一个或多个计算机处理器执行来实现一种方法的程序指令,所述方法包括:

[0008] 获得用于定义给定数据集的不同聚类的参数值,确定每个给定数据集的度量,每个给定数据集的度量取决于相应给定数据集的针对聚类之一的隶属度以及相应给定数据集到所述聚类中的同一个聚类的质心的距离,以及基于给定数据集的度量的比较从给定数据集中选择给定数据集中的至少一个以用于更新AI模块。

附图说明

[0009] 下面,仅通过示例,参考附图更详细地解释本发明的实施例,其中:

[0010] 图1描绘了用于从给定数据集中选择数据集以更新AI模块的第一计算机系统和用于执行该AI模块的第二计算机系统;

[0011] 图2描绘了包括请求输入数据集和对应的应答输出数据集的AI模块的数据流;

[0012] 图3示出了包括从图2所示的请求输入数据集和对应的应答输出数据集生成的给定数据集的日志文件;

[0013] 图4示出了串接的参数空间,其包括由串接的参数空间中的相应数据点表示的图3所示的给定数据集;以及

[0014] 图5描绘了用于从图3所示的给定数据集中选择数据集以更新AI模块的计算机实现的方法的流程图。

具体实施方式

[0015] 本发明的各种实施例的描述是为了说明的目的而呈现的,而不是旨在是穷举的或限于所公开的实施例。在不背离所描述的实施例的范围和精神的情况下,许多修改和变化对于本领域的普通技术人员将是显而易见的。选择本文所使用的术语以最好地解释实施例的原理、实际应用或对市场上存在的技术改进,或使本领域的其他普通技术人员能够理解本文所公开的实施例。

[0016] 本方法可以使得能够根据给定数据集的度量来选择给定数据集中的至少一个(以下称为所选数据集)来更新AI模块。如上所述,每个给定数据集的度量可以取决于相应给定数据集的对于聚类之一(以下称为所选聚类)的隶属度,以及相应给定数据集到所述聚类中的同一个聚类的质心(例如到所选聚类的质心)的距离。

[0017] 给定数据集的输入数据集可以具有n维,并且给定数据集的输出数据集可以具有k维。输入数据集的n维可以跨越输入参数空间,而输出数据集的k维可以跨越输出参数空间。输入数据集的n维和输出数据集的k维一起可以跨越串接的参数空间。输入参数空间、输出参数空间和/或串接参数空间可以各自具有至少一个边界。给定数据集的输入和输出数据集可以包括值,优选地包括实际值。

[0018] 给定数据集可通过使用处于训练状态的AI模块来生成。经训练的AI模块可基于对应的输入数据集之一来计算每个输出数据集。对应的输入数据集可各自表示经训练的AI模块的用户的请求,并且可被称为请求输入数据集。输出数据集可各自表示经训练的AI模块

对于对应的请求输入数据集的应答，并可被称为应答输出数据集。给定数据集可以通过将每个应答输出数据集与对应的请求输入数据集串接而被各自创建。给定数据集可由日志文件提供。当用户使用经训练的AI模块时，可通过记录应答输出数据集和对应的请求输入数据集来创建日志文件。

[0019] 给定数据集可以各自由数据点表示，数据点的坐标等于输入参数空间、输出参数空间或串接参数空间中的相应给定数据集的值，这取决于度量的计算被应用于数据集的哪个部分。短语“示例性数据集到示例性质心的示例性距离”是指示例性数据集表示的示例性数据点到示例性质心的示例性距离。类似地，短语“示例性数据集被定位到示例性质心”是指示例性数据点被定位到示例性质心，其中示例性数据集可以表示示例性数据点。

[0020] 每个给定数据集的针对所选聚类的隶属度可以基于每个给定数据集到所选聚类的质心的距离以及相应的给定数据集到除了所选聚类之外的不同聚类的质心的另外距离来确定。例如，可以基于相应给定数据集到所选聚类的质心的距离与所述另外距离和相应给定数据集到所选聚类的质心的距离之和之间的比率来确定每个给定数据集的针对所选聚类的隶属度。

[0021] 所选聚类可以从给定数据集的不同聚类中的至少两个中选择。用于定义聚类的参数值可以包括定义该聚类的每个聚类的参数值。每个聚类的参数值可以是位于输入参数空间、输出参数空间或串接参数空间中的每个聚类的质心的坐标值。所选聚类可以由与给定数据集相关的应用领域中的专家(例如医师的工程师)手动选择。在一个示例中，用于定义聚类的参数值可以通过执行应用于给定数据集、训练数据集和/或测试数据集的聚类算法来获得。在另一个示例中，可以从存储设备加载用于定义聚类的参数值。在这种情况下，可以在执行本发明的方法之前确定用于定义聚类的参数值。

[0022] 例如，专家可以指向输入参数空间、输出参数空间或串接参数空间中的位置，从而定义所选聚类的质心的坐标值。这也可以通过可视化输入参数空间、输出参数空间或串接参数空间中的三维子空间中的两个而在更高维度中实现。

[0023] 在第一示例中，每个给定数据集的度量可以通过相应给定数据集的针对所选聚类的隶属度与相应给定数据集到所选聚类的质心的距离的乘积来计算。在该第一示例中，可以执行所选数据集的选择，使得所选数据集可以是给定数据集中具有最高度量的数据集。

[0024] 根据第一示例，并且假设所选数据集的针对所选聚类的隶属度处于平均水平，例如与十个其他给定数据集相比，所选数据集可以位于相对远离所选聚类的质心的位置。在这种情况下，所选数据集可以位于比其他十个给定数据集更靠近输入参数空间、输出参数空间和/或串接参数空间的边界的位置。这可能意味着所选数据集可以包括由十个其他给定数据集给出的信息的附加信息。为此，选择所选数据集来更新AI模块可能是有趣的。

[0025] 优选地，例如可以由专家或附加AI模块来检查所选数据集。对所选数据集的检查结果可以是对所选数据集的确认或拒绝。后一种情况可表示AI模块可能已错误地计算了所选数据集的情况。在任一情况下，所选数据集可用于更新AI模块。在后一种情况下，可以优选地由专家或附加AI模块来校正所选数据集。更新AI模块可包括重新训练AI模块，例如使用所选数据集对AI模块应用反向传播算法。由于所选数据集可包括附加信息，因此更新AI模块可有助于以AI模块的改变的参数值的形式存储该附加信息。

[0026] 在另一实施例中，更新AI模块可包括改变输入参数空间或输出参数空间的边界之

一。例如,可以考虑以下两种情况。在第一种情况下,检查的结果可以是确认。在第二种情况下,检查的结果可以是拒绝。在第一种情况下,输入参数空间的边界可以移动进一步远离所选数据集。这可以具有以下优点:AI模块可以用于位于输入参数空间的经调整边界内的新数据集。在第二种情况下,可以移动输入参数空间的边界,使得所选数据集可以位于输入参数空间的边界之外。这可以降低AI模块可能为位于输入参数空间的改变的边界之外的新输入数据集计算错误的新输出数据集的风险。

[0027] 根据第二种情况改变输入参数空间的边界可以提供位于输入参数空间的改变的边界之外的新输入数据集可能不被接受用于AI模块的应用。使用查询模块可以自动执行对位于已改变边界之外的新输入数据集的拒绝,当AI模块在使用中时,所述查询模块可以用作所有到来的输入数据集的AI模块的门。AI模块可包括查询模块。查询模块可包含具有参数的函数,所述函数类似于过滤器地工作。查询模块可通过根据输入参数空间的经改变边界调整查询模块的参数的值来调整。

[0028] 包括确认或校正所选数据集的过程在此称为标注。标注可以手动或自动地执行,优选地使用附加AI模块。如果附加AI模块不是持久可访问的,具有比AI模块更好的性能,或者与AI模块相比移动性较差,则后一种情况可能是有用的。所选数据集的校正可以包括对所选数据集的输入和/或输出数据集的值之一的校正。

[0029] 本方法可使得能够在已生成给定数据集之后基于一个或多个所选数据集来更新AI模块。由于可以根据给定数据集的度量来执行对数据集的选择,因此可以考虑给定数据集在输入、输出或串接参数空间中相对于给定数据集的至少一个聚类的至少一个质心的位置。这可允许基于最重要的一个或多个给定数据集来更新AI模块。所选数据集也可以被认为是给定数据集中包含最不同信息的数据集。结果,更新AI模块可以更快,并且可以防止AI模块的过度拟合。

[0030] 根据一个实施例,该方法还包括确定每个聚类的度量,每个聚类的度量取决于相应聚类的质心到其他聚类质心的距离,基于聚类的度量从聚类选择聚类中的至少一个,以及确定每个给定数据集的度量,每个给定数据集的度量取决于相应给定数据集的针对所选聚类的隶属度以及相应给定数据集到所选聚类的质心的距离。该实施例可以具有通过比较聚类的度量来自动确定所选聚类的优点,并且在下文中可以被称为第一实施例。

[0031] 在一个示例中,每个聚类的度量可以等于相应聚类的质心到其他聚类质心的平均距离除以聚类的质心之间的最大距离的商。在第一示例中,所选聚类可以是具有最高度量的聚类。在该示例中,包括比其他给定数据集更高的针对所选聚类的隶属度的给定数据集可以比其他给定数据集更远离所有聚类质心的平衡点。由于可以基于所选聚类来计算给定数据集的度量,因此所选数据集可能比其它给定数据集更远离所述平衡点的机会可能增加。这可以增强所选数据集可以包括与其他给定数据集不同的信息的机会。

[0032] 这可以增强前一给定数据集可以包括与后一给定数据集不同的信息的机会。计算每个给定数据集的度量取决于所选聚类。

[0033] 根据一个实施例,确定每个给定数据集的度量还包括确定每个给定数据集的度量集合,相应给定数据集的度量集合中的每个度量对应于所述聚类的子集中的一个聚类,相应给定数据集的度量集合中的每个度量取决于相应给定数据集的针对相应聚类的隶属度以及相应给定数据集到相应聚类的质心的距离,以及基于给定数据集的度量集合的比较从

给定数据集中选择给定数据集中的至少一个用于更新AI模块。在一个示例中，聚类的子集可以包括所有聚类。在另一示例中，聚类的子集可以仅包括所有聚类的一部分，聚类的子集是聚类的真子集。

[0034] 根据一个示例，可以通过计算每个度量集合的范数来比较给定数据集的度量集合。一个或多个所选数据集可以分别是具有最高范数的一个或多个所选数据集。该实施例可以是有利的，因为所选数据集可以不仅取决于一个所选聚类。因此，考虑多于一个的聚类可以使用例如k均值聚类算法或模糊C均值聚类算法的聚类算法的结果，来执行从给定数据集中的选择。

[0035] 根据一个实施例，该方法还包括根据训练数据集生成用于定义聚类的参数值，AI模块是使用训练数据集生成的。以下该实施例可以被称为第二实施例。训练数据集可以包括与给定数据集相同的结构，即，训练数据集每个均包括输入数据集和输出数据集。训练数据集的功能在下文中描述，并且可以不限于该实施例。

[0036] 这里使用的术语“模块”是指任何已知的或将来开发的硬件、软件，例如可执行程序、人工智能、模糊逻辑或其任何可能的组合，用于执行与“模块”相关联的功能或作为已经执行与“模块”相关联的功能的结果。

[0037] AI模块可包括神经网络、卷积神经网络和/或径向基函数网络。给定数据集和训练数据集的输入数据集和输出数据集可以包括作为数据元素的值，优选地是实际值。可以根据相应的输入数据集和AI模块的参数值来执行给定数据集和训练数据集的输出数据集之一的计算。在优选的示例中，给定数据集和训练数据集的每个输出数据集的值可以各自表示给定数据集和训练数据集的输入数据集分别可以被分类到几个类别的哪个类别中的概率。

[0038] AI模块可以使用机器学习基于训练数据集来生成。术语“机器学习”是指用于从训练数据集的输入数据集和输出数据集提取有用信息的计算机算法。可以通过以自动化方式建立概率模型来提取信息。可以使用一个或多个已知的机器学习算法来执行机器学习，例如线性回归、反向传播、K均值、分类算法等。

[0039] 概率模型可以例如是使得能够基于训练数据集的输入数据集之一来预测类别或者将与训练数据集的输入数据集之一相对应的实例与相应输出数据集的一个或多个值相关联的等式或规则集。

[0040] 所述一个或多个已知机器学习算法可以调整所述AI模块的参数值，使得可以减少所述AI模块的训练误差。可以基于由AI模块计算的AI模块的训练输出数据集的计算值与相应训练数据集的每个输出数据集的值的偏差来计算训练误差。AI模块的每个训练输出数据集可以基于相应训练数据集的输入数据集来计算，并且因此可以与相应训练数据集相关联。AI模块的训练输出数据集可以具有与训练数据集的输出数据集相同的结构，即AI模块的训练输出数据集的元素类型可以匹配训练数据集的输出数据集的元素类型。

[0041] 基于所述偏差来调整AI模块的参数值可以减少训练误差。如果训练误差达到给定阈值，则AI模块可被视为正在被训练并处于已训练状态。在已训练状态中，AI模块可以用于分别响应于由用户发送到AI模块的请求输入数据集而生成上述应答输出数据集。

[0042] 训练数据集可被选择成使得训练数据集的输入数据集可在输入参数空间中尽可能相等地分布和/或使得它们可表示AI模块可被应用于的许多重要用例。训练数据集的分

布可以被设计成使得训练误差可以尽可能低。这可能意味着在串接参数空间的不同区域中，训练数据集的密度可以不同。可以使用实验设计(DOE)算法来计算串接参数空间中的训练数据集的推荐的不同密度。不同的密度可以被认为是训练聚类。

[0043] 通常，训练数据集可以以监督方式获得，例如通过考虑推荐密度来获得它们，通过在监督的和/或设计的实验中获得它们，和/或通过从一组实验数据集中选择训练数据集。这种监督可以由专家执行。为此，训练数据集可以比给定数据集更高效地表示专家的知识。例如，给定数据集可通过在串接参数空间的非常窄的子空间中使用AI模块来生成，该子空间仅覆盖AI模块的非常少的不同用例。

[0044] 根据训练数据集生成用于定义聚类的参数值可以提供：聚类可以被专家容易地理解并且可以表示串接参数空间的有意义的聚类。此外，聚类可以反映输入、输出或串接参数空间中训练数据集的不同密度。此外，与仅使用给定数据集进行聚类相比，可以更快地执行聚类算法。因此，在优选实施例中，可以仅使用训练数据集来生成用于定义聚类的参数值。

[0045] 根据一个实施例，该方法还包括根据给定数据集生成用于定义聚类的参数值。以下，该实施例可以被称为第三实施例。给定数据集可表示可能不被训练数据集包括的AI模块的新用例。因此，使用给定数据集进行聚类所产生的聚类可以表示包含新用例的输入、输出或串接参数空间的新区域。所选数据集可以位于新区域之一中并且表示新用例之一。因此，AI模块可使用包括由新用例之一表示的新信息的所选数据集来更新。

[0046] 根据一个实施例，该方法还包括根据测试数据集生成用于定义聚类的参数值，AI模块是使用测试数据集来测试的。以下，该实施例可以被称为第四实施例。测试数据集可以具有与训练数据集相同的结构，即，每个测试数据集包括输入和输出数据集。测试数据集可以源自实验数据集的集合，并且因此可以以类似于训练数据集的方式表示专家的知识。由于该原因，该实施例可以具有与仅使用训练数据集进行聚类相同的优点。如果用于定义聚类的参数值是根据测试数据集和训练数据集而生成的，则可以使用更多的信息，并且聚类可以更好地表示专家的知识。测试数据集可用于AI模块的验证。该验证可以在下面描述。

[0047] 可以基于由AI模块计算的AI模块的验证输出数据集的计算值与相应测试数据集的每个输出数据集的值的偏差来计算验证误差。AI模块的每个验证输出数据集可基于相应测试数据集的输入数据集来计算，并因此可与相应测试数据集相关联。AI模块的验证输出数据集可具有与测试数据集的输出数据集相同的结构，即AI模块的验证输出数据集的元素的类型可匹配测试数据集的输出数据集的元素的类型。

[0048] 如果验证误差达到给定的验证阈值，则AI模块可被认为是已验证的。如果验证误差与验证阈值不匹配，则机器学习算法之一可以被重复地执行，以便再次适配AI模块的参数值。在这种情况下，AI模块的参数值可以被不同地初始化。如果AI模块被验证，则它可以提供足够的一般化属性，即，在新输入数据集的基础上计算足够准确的新输出数据集。

[0049] 根据一个实施例，该方法还包括根据给定数据集中的批准或校正的数据集(在下文中称为标注数据集)生成用于定义聚类的参数值。对要标注的给定数据集之一的批准或校正(即，标注)可以由专家手动执行，或者例如由附加AI模块自动执行。批准或校正可以包括对待标注的一个数据集的输入数据集和/或输出数据集的批准或校正。当输入数据集的值可能已知是错误的(例如移位了已知值)时，校正该输入数据集可能是合理的。可以完成校正输出数据集以校正AI模块的预测。根据标注数据集生成用于定义聚类的参数值可能是

有利的，因为聚类可以基于标注数据集所包括的新信息来执行。

[0050] 根据一个实施例，该方法还包括根据给定数据集中的手动批准或手动校正的数据集来生成用于定义聚类的参数值。以下，本实施例可以被称为第五实施例。在该实施例中，标注数据集可以手动创建，例如由上述专家创建，并且由此可以更可靠和透明地创建。

[0051] 根据一个实施例，该方法还包括获得用于定义执行模糊C均值聚类算法的聚类的参数值。以下，该实施例可以被称为第六实施例。模糊C均值聚类算法可以应用于给定数据集、训练数据集和/或测试数据集。与使用另一种聚类算法(例如k均值聚类算法)相比，使用模糊C均值聚类算法的优点可能是，聚类的解可以更少地依赖于聚类的质心的初始选择。这可以导致聚类的更一致的解。与k均值聚类算法相比，执行模糊C均值聚类算法可以包括将每个给定数据集的隶属度分配给每个聚类。可以给出多个聚类用于执行模糊C均值聚类算法。

[0052] 根据一个实施例，该方法还包括基于训练数据集的输入数据集获得用于定义聚类的参数值。优选地，可以仅基于给定数据集、训练数据集和/或测试数据集的输入数据集来执行聚类。这可能是有利的，因为聚类的解可能不取决于AI模块的准确性。这可以允许由专家解释该解决方案而不那么混淆。

[0053] 根据一个实施例，该方法还包括基于训练数据集的输出数据集获得用于定义聚类的参数值。优选地，可以仅基于给定数据集、训练数据集和/或测试数据集的输出数据集来执行聚类。通常，给定数据集或训练数据集的每个输出数据集的值的数量小于给定数据集或训练数据集的对应输入数据集的值的数量。在这种情况下，该实施例可以减少聚类的数量。在这种情况下，聚类的解可能更容易理解。此外，更新AI模块以使得可以减少对可以由给定数据集或训练数据集的输出数据集表示的若干类中的一个类的预测误差可能是有用的。在这种情况下，仅对给定数据集、训练数据集和/或测试数据集的输出数据集的聚类可以更高效。若干类中的同一个类可以由聚类之一来表示。该聚类可以被手动选择为用于选择给定数据集中的至少一个的所选聚类。

[0054] 根据一个实施例，该方法还包括基于训练数据集的输入数据集和输出数据集来获得用于定义聚类的参数值。优选地，可以基于给定数据集、训练数据集和/或测试数据集的输出数据集和输入数据集来执行聚类。该实施例可以导致表示尽可能多的给定数据集、训练数据集和/或测试数据集的信息的聚类。

[0055] 参考最后三个实施例，如果仅给定数据集、训练数据集和/或测试数据集的输入数据集用于聚类，则可以仅基于给定数据集的输入数据集来计算给定数据集的度量。类似地，如果仅给定数据集、训练数据集和/或测试数据集的输出数据集用于聚类，则可以仅基于给定数据集的输出数据集来计算给定数据集的度量。以相同的方式，如果给定数据集、训练数据集和/或测试数据集的输入数据集和输出数据集被用于聚类，则可以基于给定数据集的输出数据集和输入数据集来计算给定数据集的度量。

[0056] 根据一个实施例，该方法还包括基于给定数据集到相应聚类的质心的平均距离来确定每个聚类的度量。以下，该实施例可以被称为第七实施例。在优选实施例中，可以计算每个聚类的度量，使得给定数据集到相应聚类的质心的平均距离的较高值可以引起相应聚类的度量的较低值。在这种情况下，如果具有最低度量的聚类是所选聚类，则所选聚类可以是其中给定数据集在相应聚类内更分散的聚类之一。

[0057] 根据一个实施例，该方法还包括基于给定数据集到相应聚类的质心的最大距离来确定每个聚类的度量。该实施例在下文中可以被称为第八实施例。在优选实施例中，可以计算每个聚类的度量，使得给定数据集到相应聚类的质心的平均距离的较高值可以引起相应聚类的度量的较高值。在这种情况下，如果具有最低度量的聚类是所选聚类，则远离相应聚类的质心定位的给定数据集的离群值可以指示该聚类不是所选聚类。因此，该实施例防止了给定数据集的离群值可能对所选聚类的确定具有强烈影响。如果给定数据集到质心的最大距离和给定数据集到质心的平均距离以上述方式一起用于确定所选聚类，则该实施例可以指示给定数据集的离群值对平均距离的值的影响可以通过它们对最大距离的影响来平衡。

[0058] 根据一个实施例，该方法还包括基于给定数据集到相应聚类的平均隶属度来确定每个聚类的度量。该实施例在下文中可以被称为第九实施例。优选地，可以基于给定数据集和训练数据集对于相应聚类的平均隶属度来确定每个聚类的度量。在优选实施例中，可以计算每个聚类的度量，使得给定数据集和/或训练数据集对于相应聚类的平均隶属度的较高值可以导致相应聚类的度量的较高值。在这种情况下，如果具有最低度量的聚类是所选聚类，则所选聚类可以是包括更多给定数据集的聚类之一，其中给定数据集对于相应聚类的隶属度更低。因此，所选聚类可以包括不太清楚或不太容易被分类的相应的给定数据集。如果所选数据集源于以这种方式确定的所选聚类，则所选数据集可能包括新信息的机会可以增加。

[0059] 在第七、第八和第九实施例中，可以优选地根据训练数据集和给定数据集的手动批准或手动较正的数据集来生成用于定义聚类的参数值。可以响应于给定数据集的扩展而重复根据第七、第八和第九实施例的步骤。给定数据集可在AI模块的使用期间扩展。在此使用期间，可扩展日志文件，使得新的给定数据集可由日志文件包括。如果在执行根据第七、第八和第九实施例的步骤的第一次迭代中没有给定数据集的手动标注的数据集，则用于定义聚类的参数值可以优选地仅根据训练数据集而生成。

[0060] 根据一个实施例，该方法还包括基于训练数据集和给定数据集的手动批准或手动校正的数据集到相应聚类的质心的平均距离来确定每个聚类的度量。以下，该实施例可以被称为第十实施例。

[0061] 根据一个实施例，该方法还包括基于训练数据集和给定数据集的手动批准或手动校正的数据集到相应聚类的质心的最大距离来确定每个聚类的度量。以下，该实施例可以被称为第十一实施例。

[0062] 根据一个实施例，该方法还包括基于训练数据集和给定数据集的手动批准或手动校正的数据集针对相应聚类的平均隶属度来确定每个聚类的度量。以下，该实施例可以被称为第十二实施例。

[0063] 第十、第十一和第十二实施例可具有与第七、第八和第九实施例相似的优点。基于训练数据集和给定数据集的手动批准或手动较正的数据集来确定每个聚类的度量可以具有以下优点：所选聚类可以仅基于批准和手动校正的数据集来确定。结果，专家可以容易地确定聚类的选择。然而，基于给定数据集确定每个聚类的度量可以增加所选聚类可能包括由所选数据集提供的新信息的机会。

[0064] 根据一个实施例，该方法还包括基于相应聚类所包括的训练数据集的数量与相应

聚类所包括的给定数据集的手动批准或手动校正的数据集的数量的第一和与训练数据集的总数与给定数据集的手动批准或手动校正的数据集的总数的第二和的比率，来确定每个聚类的度量。该实施例在下文中可以被称为第十三实施例。在优选实施例中，可以计算每个聚类的度量，使得该比率的较高值可以引起相应聚类的度量的较高值。在这种情况下，如果具有最低度量的聚类是所选聚类，则所选聚类可以是包括更少手动标注的数据集和训练数据集的聚类之一。因此，所选聚类可以包括低密度的数据集。

[0065] 在第十、第十一、第十二和第十三实施例中，可以优选地根据训练数据集、测试数据集以及给定数据集的手动批准或手动校正的数据集来生成用于定义聚类的参数值。

[0066] 类似于根据第七、第八和第九实施例的步骤，可以响应于给定数据集的扩展而重复第十、第十一、第十二和第十三实施例的步骤。如果在执行根据第十、第十一、第十二和第十三实施例的步骤的第一迭代中没有给定数据集的手动标注的数据集，则可以优选地仅根据训练数据集和测试数据集来生成用于定义聚类的参数值。

[0067] 根据一个实施例，给定数据集的输入数据集均包括标识参数的值，并且给定数据集的输出数据集均包括性能指标的值。在该实施例中，输出参数空间可以包括性能指标，而输入参数空间可以包括标识参数。这可以使得能够根据给定数据集的性能指标和/或标识参数的每个值来确定所选数据集。此外，该实施例可以使得能够根据性能指标的值来更新AI模块。

[0068] 标识参数可以允许将每个给定数据集与数据处理的相应动作相关联。数据处理的相应动作可以包括相应给定数据集的生成。例如，考虑日志文件，相应给定数据集的所述标识参数可以是与串接相应给定数据集的输入数据集与相应给定数据集的输出数据集并将它们以相应给定数据集的形式写入日志文件的实例相关的识别编号。在此示例中，每当日志文件由另一给定数据集扩展时，可增加识别编号。

[0069] 相应给定数据集的输入数据集可以包括第一另外值，其可以与相应给定数据集的生成实例相关，优选地与相应给定数据集的输出数据集的生成实例相关。该输入数据集的第一另外值可以包括关于影响输出数据集的值的环境状态的信息，优选地包括相应给定数据集的性能指标的值。在另一实施例中，可以使用第一另外值来计算标识参数的值，该第一另外值可以与相应给定数据集的生成实例相关，优选地与相应给定数据集的输出数据集的生成实例相关。标识参数的值可以由第一函数计算，该第一函数可以将第一另外值的组合双射地映射到标识参数的值。

[0070] 性能指标的值可以与通信的性能相关。例如，如果通信成功，则性能指标的值在替代方案中可以等于一和零。该通信可以与相应给定数据集的第二另外值相关。第二另外值可以指定动作，例如通信。例如，可以通过指示相应给定数据集的输入数据集已经被发送到哪个目的地、相应给定数据集的输入数据集包括哪种信息、和/或相应给定数据集的输入数据集的发送可能引起哪种动作，来指定通信。第二另外值可以包含在相应给定数据集的输入和/或输出数据集中。

[0071] 图1示出了用于从给定数据集14(图3中所示)中选择数据集以更新人工智能模块(AI模块)1(图2中所示)的第一计算机系统100。第一计算机系统100可以适于执行根据本发明的各种实施例的方法步骤。第一计算机系统100可以包括通过第一总线106耦接在一起的第一处理器102、第一存储器103、第一I/O电路104和第一网络接口105。

[0072] 第一处理器102可以表示一个或多个处理器(例如微处理器)。第一存储器103可以包括易失性存储器元件(例如,随机存取存储器(RAM,诸如DRAM、SRAM、SDRAM等))和非易失性存储器元件(例如,ROM、可擦除可编程只读存储器(EPROM)、电可擦除可编程只读存储器(EEPROM)和可编程只读存储器(PROM))中的任何一个或组合。要注意,第一存储器103可以具有分布式结构,其中各种部件彼此远离地定位,但是可以由第一处理器102访问。

[0073] 第一存储器103与第一持久性存储设备107组合可以用于本地数据和指令存储。第一存储设备107包括由第一I/O电路104控制的一个或多个持久性存储设备和介质。第一存储设备107可以包括用于数字数据存储的磁、光、磁光或固态装置,例如具有固定或可移动介质。样本设备包括硬盘驱动器、光盘驱动器和软盘驱动器。样本介质包括硬盘盘片、CD-ROM、DVD-ROM、BD-ROM、软盘等。

[0074] 第一存储器103可以包括一个或多个单独的程序,每个程序包括用于实现逻辑功能、特别是在示例中涉及的功能的可执行指令。第一存储器103中的软件通常还可以包括适当的第一操作系统(OS)108。第一OS108实质上控制用于实现如本文所述的方法的至少一部分的其它计算机程序的执行。

[0075] 第一计算机系统100可以被配置为获得用于定义给定数据集14的不同聚类的参数值,在下文中被称为第一功能。第一功能可以包括加载第一值和第二值,第一值可以指示不同聚类的质心的坐标,第二值可以指示每个给定数据集的针对每个聚类的隶属度。第一功能可以包括使用给定数据集14、训练数据集和/或测试数据集来执行聚类算法,例如模糊C均值聚类算法。

[0076] 此外,第一计算机系统100可以被配置为确定每个给定数据集的度量,每个给定数据集的度量取决于相应给定数据集的针对聚类之一的每个隶属度以及相应给定数据集到所述聚类中的同一聚类的质心的距离,在下文中被称为第二功能。

[0077] 此外,第一计算机系统100可被配置用于诸如基于给定数据集14的度量的比较从给定数据集14中选择给定数据集14中的至少一个来更新AI模块1(图2中所描绘)的功能,在下文中被称为第三功能。

[0078] 此外,第一计算机系统100可以被配置为确定每个聚类的度量,每个聚类的度量取决于相应聚类的质心到其他聚类质心的距离,并且基于聚类的度量从所述聚类中选择所述聚类中至少一个,在下文中被称为第四功能。每个给定数据集的度量可以根据上述方法之一进行计算。

[0079] 此外,第一计算机系统100可以被配置成生成用于根据第二、第三、第四、第五和第六实施例定义聚类的参数值,以下分别称为第五、第六、第七、第八和第九功能。

[0080] 此外,第一计算机系统100可以被配置成根据第七、第八、第九、第十、第十一、第十二和第十三实施例确定每个聚类的度量,在下文中分别被称为第十、第十一、第十二、第十三、第十四、第十五和第十六功能。

[0081] 第一计算机系统100可以通过分别执行第一程序201、第二程序202、第三程序203、第四程序204、第五程序205、第六程序206、第七程序207、第八程序208、第九程序209、第十程序210、第十一程序211、第十二程序212、第十三程序213、第十四程序214、第十五程序215和第十六程序216,来执行第一、第二、第三、第四、第五、第六、第七、第八、第九、第十、第十一、第十二、第十三、第十四、第十五和第十六功能。第一处理器102可以执行主程序200。根

据确定用于定义聚类的参数值和每个聚类的度量的特定实施例，主程序200可以启动程序201、202、203、204、205、206、207、208、209、210、211、212、213、214、215和216在第一处理器102上的执行。

[0082] 如本文所使用的术语“程序”是指一组指令，其包括当处理器102可以读取命令时用于引起由处理器102执行的动作的命令。该组指令可以是计算机可读程序、例程、子例程或库的一部分的形式，其可以由处理器102执行和/或可以由处理器102执行的另外的程序调用。优选地，程序200、201、202、203、204、205、206、207、208、209、210、211、212、213、214、215、216可以是根据计算机系统100的硬件平台的类型编译的可执行程序。第一存储器103可以包括用于存储程序200、201、202、203、204、205、206、207、208、209、210、211、212、213、214、215、216的空间，该空间在下文中被称为第一功能存储器115。

[0083] 图1示出了第二计算机系统120。第二计算机系统120可以适于执行AI模块1(图2中所示)。

[0084] 第二计算机系统120可以包括通过第二总线126耦接在一起的第二处理器122、第二存储器123、第二I/O电路124和可以被设计为第二网络接口的网络接口2。

[0085] 第二处理器122可以表示一个或多个处理器(例如微处理器)。第二存储器123可以包括易失性存储器元件(例如，随机存取存储器(RAM，诸如DRAM、SRAM、SDRAM等))和非易失性存储器元件(例如，ROM、可擦除可编程只读存储器(EPROM)、电可擦除可编程只读存储器(EEPROM)、可编程只读存储器(PROM))中的任何一个或组合。要注意，第二存储器123可以具有分布式架构，其中各种组件彼此远离地定位，但是可以由第二处理器122访问。

[0086] 第二存储器123与第二持久性存储设备127组合可以用于本地数据和指令存储。第二存储设备127包括由第二I/O电路124控制的一个或多个持久性存储设备和介质。第二存储设备127可以包括用于数字数据存储的磁、光、磁光或固态装置，例如具有固定或可移动介质。样本设备包括硬盘驱动器、光盘驱动器和软盘驱动器。样本介质包括硬盘盘片、CD-ROM、DVD-ROM、BD-ROM、软盘等。

[0087] 第二存储器123可以包括一个或多个单独的程序，每个程序包括用于实现逻辑功能、特别是在示例中涉及的功能的可执行指令。第二存储器123中的软件通常还可以包括适当的第二操作系统(OS)128。第二OS 128实质上控制用于实现如本文所述的方法的至少一部分的其它计算机程序的执行。

[0088] 第二计算机系统120可以被配置成在第二计算机系统120上执行AI模块1(图2中描绘)，这在下面被称为第十七功能。第十七功能可以包括将神经网络、卷积神经网络和/或径向基函数网络的模型函数的参数值和结构从第二存储设备127加载到第二存储器123中，并且基于相应的请求输入数据集来计算应答输出数据集。基于其可以计算应答输出数据集的请求输入数据集可以对应于该应答输出数据集，反之亦然。

[0089] 如图2所示，AI模块1可以类似于应答输出数据集计算一组应答输出数据集10，其中每个应答输出数据集可以基于一组请求输入数据集9中的单个对应的请求输入数据集来计算。

[0090] 此外，第二计算机系统120可以被配置为经由接口2接收请求输入数据集9，在下文中被称为第十八功能，并且经由接口2发送应答输出数据集10，在下文中被称为第十九功能。

[0091] 第二计算机系统120可以通过分别执行第十七程序217、第十八程序218和第十九程序219来执行第十七、十八和十九功能。程序217、218、219的执行可以通过在第二处理器122上执行第二主程序220来启动。第二存储器123可以包括用于存储程序220、217、218、219的空间，该空间在下文中被称为第二功能存储器135。

[0092] AI模块1(图2中所示)可被认为是包括模型函数的参数值和结构的实体，并且用于在第二处理器122上运行神经网络、卷积神经网络和/或径向基函数网络的程序217被加载到第二处理器122的高速缓存中。

[0093] 给定数据集14(图3中所示)中的每一个可以通过将应答输出数据集10(图2中所示)中的一个与请求输入数据集9(图2中所示)中的对应一个串接而创建。优选地，给定数据集14中的每一个可以被划分成输入数据集和输出数据集。请求输入数据集9中的每一个可以包括与给定数据集14的输入数据集11(图3中描绘)之一相同的值，并且应答输出数据集10中的每一个可以与给定数据集14的输出数据集12(图3中描绘)之一相同。因此，在该示例中，当根据请求输入数据集9和应答输出数据集10创建给定数据集14时，请求输入数据集9可以成为给定数据集14的输入数据集11，并且应答输出数据集可以成为给定数据集14的输出数据集12。

[0094] 给定数据集14可由图3所示的日志文件13提供。日志文件13可通过在用户使用经训练的AI模块1时存储应答输出数据集12和对应的请求输入数据集11来创建。优选地，每当AI模块1计算新的应答输出数据集时，日志文件13可被扩展另一给定数据集。在一个示例中，日志文件13可由第二计算机系统120创建且存储在第二存储器123中。在另一个示例中，日志文件13可以由第一计算机系统100创建，优选地通过分别读入请求输入数据集11和应答输出数据集12来创建。

[0095] 在一个示例中，AI模块1可以在第一处理器102上执行。然而，本发明的实施例可在不访问AI模块1的情况下执行。由于这可能更经常发生，因此在图1和2中描述了该示例。可以仅需要给定数据集14来执行本发明的实施例。优选地，通过加载日志文件13，给定数据集可被加载到第一存储器103中。为了实现这一点，第一网络接口105可经由万维网130或另一网络与接口2通信地耦接。

[0096] 在一个示例中，输入数据集11可以各自包括如图3中的a_1、a_i、a_n所示的第一值和如图3中的b_1、b_i、b_n所示的第二值，并且输出数据集12可以各自包括如图3中的c_1、c_i、c_n所示的第一值。

[0097] 给定数据集14可以各自由坐标系40(图4中描绘)中的数据点表示，其中每个数据点的坐标等于相应给定数据集的值。图4示出了一些示例性数据点41，通过这些数据点可以表示给定数据集14。在这种情况下，坐标系40可以表示包括给定数据集14的输入参数空间和输出参数空间的串接参数空间。给定数据集14的输入参数空间可以跨越x轴42和y轴43，并且可以包括输入数据集11的第一值a_1、a_i、a_n和第二值b_1、b_i、b_n。给定数据集14的输出参数空间可以跨越z轴44，并且可以包括输出数据集12的第一值c_1、c_i、c_n。

[0098] AI模块1可以处于用于执行本方法的已训练状态。在AI模块1的未训练状态中，模型函数的参数值可以等于随机值。这可以通过AI模块1的初始化来实现，其中模型函数的参数值可以被设置为随机值。AI模块1的训练可以基于训练数据集46(图4中描绘)来执行，每个训练数据集46包括输入数据集和输出数据集。

[0099] 训练数据集46的输入和输出数据集可以具有元素。这些元素可以是值，优选地是实际值。训练数据集46的输入数据集可以具有与给定数据集14的输入数据集11相同的结构。类似地，训练数据集46的输出数据集可以具有与给定数据集14的输出数据集12相同的结构。训练数据集46可以表示关于分类问题的信息，一旦用训练数据集46进行了训练，AI模块1就可以用于该分类问题。关于第一用例，各个输入数据集11的第一值a_1、a_i、a_n和第二值b_1、b_i、b_n可以各自是用于将相应输入数据集11分组为若干不同类之一的特征值。每个不同类的类型可以由相应输出数据集12的第一值c_1、c_i、c_n给出。训练数据集46的每个输入和输出数据集的值可以具有与给定数据集14相同的结构，并且可以通过实验获得，优选地通过受监督实验获得。

[0100] 可以执行AI模块1的训练，使得可以调整模型函数的参数值，以减少AI模块1的训练误差。训练误差可以如上所述使用一个或多个学习算法来减小，所述学习算法例如为线性回归、反向传播、K均值等。

[0101] 图5示出了用于从给定数据集14中选择数据集以更新AI模块1的计算机实现的方法的流程图，每个给定数据集14_i(图3中描绘)包括输入数据集11_i(图3中描绘)和对应的输出数据集12_i(图3中描绘)。

[0102] 在步骤301中，可以获得用于定义给定数据集14的不同聚类45的参数值。这可以通过在第一处理器102上执行第一程序201来实现。运行第一程序201，可以基于训练数据集46执行模糊C均值聚类算法。这可以包括确定聚类45的质心47(图4中所示)以及给定数据集14_i中的每一个的针对聚类45中的每一个的隶属度。

[0103] 在步骤302中，可以确定每个给定数据集14_i的度量。每个给定数据集的度量可以取决于相应给定数据集14_i的针对聚类之一的隶属度以及相应给定数据集14_i到所述聚类中的同一个聚类的质心的距离。

[0104] 在步骤303中，可以基于给定数据集14的度量的比较从给定数据集14中选择给定数据集14中的至少一个用于更新AI模块1。

[0105] 在第一示例中，可以确定聚类45中的每一个的度量。聚类45的每个聚类的度量可以取决于聚类45的相应聚类的质心到聚类45的其它质心的距离。此外，可以基于聚类45的度量从聚类45中选择聚类45中的一个聚类。根据该第一示例，可以确定每个给定数据集14_i的度量，使得每个给定数据集14_i的度量可以取决于相应给定数据集14_i的针对所选聚类的隶属度以及相应给定数据集14_i到所选聚类的质心的距离。相应给定数据集14_i到所选聚类的质心的距离可以等于相应数据点到所选聚类的质心的距离，该相应数据点可以表示相应给定数据集14_i。

[0106] 例如，每个给定数据集14_i的度量$Mdat_i$可以根据下式计算：

[0107] $$Mdat_i = \frac{1}{2}\left(M + \frac{D}{MD}\right)$$

[0108] 其中，D可以是相应给定数据集14_i到所选聚类的质心的距离，MD可以是给定数据集14到所选聚类的质心的最大距离，并且M是相应给定数据集14_i的针对所选聚类的隶属度。

[0109] 根据第一示例的第一变型，聚类45中的每一个(在下面被称为聚类45_i)的度量$Mclust1_i$可以根据下式来确定：

[0110] $$Mclust1_i = \frac{1}{4}\left(R + \left(1 - \frac{MeanD1}{MaxD1}\right) + MM1 + MCD1\right)$$

[0111] 其中，MeanD1可以是训练数据集46到相应聚类45_i的质心的平均距离或者训练数据集46和标注数据集到相应聚类45_i的质心的平均距离。标注数据集可以每个是给定数据集14的批准或校正数据集。如上所述，对给定数据集之一14_i的批准或校正(即，标注)可以由专家手动执行或者可以自动执行。

[0112] 此外，MM1可以是训练数据集46的针对相应聚类45_i的隶属度的平均值，或者是训练数据集46和标注数据集的针对相应聚类45_i的隶属度的平均值。此外，MaxD1可以是训练数据集46到聚类45的质心的最大距离或者训练数据集46和标注数据集到聚类45的质心的最大距离。此外，MCD1可以是从相应聚类45_i的质心到其它聚类45的平均距离除以到聚类45的质心的最大距离。此外，R可以是由相应聚类45_i包括的标注数据集和训练数据集46的第一和与所有训练数据集46和所有标注数据集的第二和的比率。

[0113] 根据第一示例的第二变型来确定聚类45中的每一个的度量$Mclust1_i$，组合了上述第十、第十一、第十二和第十三实施例，并且可产生针对这些实施例所述的优点。程序213、214、215和216可以在第一处理器102上执行以确定聚类45中的每一个的度量$Mclust1_i$，并且可以由主程序200调用。

[0114] 根据第一示例的第一变型，所选聚类可以是包括度量$Mclust1_i$的最低值的聚类。用于获得聚类45的质心和每个给定数据集14_i的针对聚类45中的每一个的隶属度值的聚类可以基于训练数据集46、上述测试数据集、给定数据集14和/或标注数据集来执行。在这种情况下，训练数据集46、上述测试数据集、给定数据集14和/或标注数据集可以构建可以对其执行聚类的一组数据集。

[0115] 根据第一示例的第二变型，可根据下式确定聚类45中的每一个的度量$Mclust2_i$：

[0116] $$Mclust2_i = \frac{1}{3}\left(\left(1 - \frac{MeanD2}{MaxD2}\right) + MM2 + MCD2\right)$$

[0117] 其中，MeanD2可以是给定数据集14到相应聚类45_i的质心的平均距离。此外，MM2可以是给定数据集14的针对相应聚类45_i的隶属度的平均值。此外，MaxD2可以是给定数据集14到聚类45的质心的最大距离。此外，MCD2可以是从相应聚类45_i的质心到其他聚类45的平均距离除以到聚类45的质心的最大距离。

[0118] 根据第一示例的第一变型来确定聚类45中的每一个的度量$Mclust2_i$，组合了上述第七、第八和第九实施例，并且可产生针对这些实施例所述的优点。程序210、211和212可以在第一处理器102上执行以确定聚类45中的每一个的度量$Mclust2_i$，并且可以由主程序200调用。

[0119] 根据第一示例的第二变型，所选聚类可以是包括度量$Mclust2_i$的最低值的聚类。用于获得聚类45的质心和每个给定数据集14i的针对聚类45中的每一个的隶属度值的聚类可以基于训练数据集46和/或标注数据集来执行。在这种情况下，训练数据集46和/或标注数据集可以构建可以对其执行聚类的一组数据集。

[0120] 下面描述了如何基于每个给定数据集14_i的度量$Mdat_i$的比较从给定数据集14中选择多于一个的数据集。在这种情况下，可以根据第一示例的第一或第二变型来确定所选聚类。度量$Mdat_i$的最小值Min_Mdat_i和度量$Mdat_i$的最大值Max_Mdat_i可以通过比较给定数据

集14_i的度量$Mdat_i$来确定。包括最小值Min_Mdat_i和最大值Max_Mdat_i作为其边界的范围可以被划分为N个相等的子范围，每个子范围包括最小和最大边界。给定数据集14_i可以根据它们的度量$Mdat_i$以及所述N个子范围的最小边界值和最大边界值被分组成N个不同的组。从Min_Mdat_i的给定数据集。从N个不同的组中的每个组，可以选择给定数量M个给定数据集。从不同的组中的每个组选择给定数量M个数据集可具有从给定数据集14创建关于所选聚类的所选数据集的同构组的优点。

[0121] 根据另一示例，所有给定数据集14_i中包括最低度量$Mdat_i$的数据集可以被选择，或者所有给定数据集14_i中包括最低度量$Mdat_i$的给定数量L个数据集可以被选择。在另一示例中，所有给定数据集14_i中包括最高度量$Mdat_i$的数据集可以被选择，或者所有给定数据集14_i中包括最高度量$Mdat_i$的给定数量L个数据集可以被选择。

[0122] 与选择一个或多个数据集的方法无关，可以手动或自动地标注一个或多个所选数据集，以便生成上述的一个或多个标注数据集。基于一个或多个标注数据集，响应于用新的给定数据集扩展日志文件13，可以如上所述执行聚类。上面描述了可以如何创建新的给定数据集。

[0123] 当AI模块1在使用中时，选择一个或多个数据集并分别标注一个或多个数据集的所述过程可以重复执行，重复创建新的给定数据集，并由此扩展日志文件13，并由此增加给定数据集14的数量。标注数据集可以用于更新AI模块1。更新可以以类似于上述AI模块1的训练的重新训练的形式执行，但至少基于标注数据集。重新训练也可以基于训练数据集和标注数据集来执行。

[0124] 本发明可以是任何可能的技术细节集成水平的系统、方法和/或计算机程序产品。计算机程序产品可以包括其上具有计算机可读程序指令的计算机可读存储介质，所述计算机可读程序指令用于使处理器执行本发明的各方面。计算机可读存储介质可以是能够保留和存储由指令执行设备使用的指令的有形设备。

[0125] 计算机可读存储介质可以是例如但不限于电子存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或前述的任何合适的组合。计算机可读存储介质的更具体示例的非穷举列表包括以下：便携式计算机磁盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式光盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、诸如上面记录有指令的打孔卡或凹槽中的凸起结构的机械编码装置，以及上述的任何适当组合。如本文所使用的计算机可读存储介质不应被解释为暂时性信号本身，诸如无线电波或其他自由传播的电磁波、通过波导或其他传输介质传播的电磁波(例如，通过光纤线缆的光脉冲)、或通过导线传输的电信号。

[0126] 本文描述的计算机可读程序指令可以从计算机可读存储介质下载到相应的计算/处理设备，或者经由网络，例如因特网、局域网、广域网和/或无线网络，下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光传输光纤、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或网络接口从网络接收计算机可读程序指令，并转发计算机可读程序指令以存储在相应计算/处理设备内的计算机可读存储介质中。

[0127] 用于执行本发明的操作的计算机可读程序指令可以是汇编指令、指令集架构

(ISA)指令、机器相关指令、微代码、固件指令、状态设置数据、集成电路的配置数据，或者以一种或多种编程语言(包括面向对象的编程语言，例如Smalltalk、C++等)和过程编程语言(例如“C”编程语言或类似的编程语言)的任意组合编写的源代码或目标代码。计算机可读程序指令可以完全在用户的计算机上执行，部分在用户的计算机上执行，作为独立的软件包执行，部分在用户的计算机上并且部分在远程计算机上执行，或者完全在远程计算机或服务器上执行。在后一种情况下，远程计算机可以通过任何类型的网络连接到用户的计算机，包括局域网(LAN)或广域网(WAN)，或者可以连接到外部计算机(例如，使用因特网服务提供商通过因特网)。在一些实施例中，为了执行本发明的各方面，包括例如可编程逻辑电路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA)的电子电路可以通过利用计算机可读程序指令的状态信息来执行计算机可读程序指令以使电子电路个性化。

[0128] 在此参考根据本发明实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述本发明的各方面。将理解，流程图和/或框图的每个框以及流程图和/或框图中的框的组合可以由计算机可读程序指令来实现。

[0129] 这些计算机可读程序指令可以被提供给计算机或其他可编程数据处理装置的处理器以产生机器，使得经由计算机或其他可编程数据处理装置的处理器执行的指令创建用于实现流程图和/或框图的一个或多个框中指定的功能/动作的装置。这些计算机可读程序指令还可以存储在计算机可读存储介质中，其可以引导计算机、可编程数据处理装置和/或其他设备以特定方式工作，使得其中存储有指令的计算机可读存储介质包括制品，该制品包括实现流程图和/或框图的一个或多个框中指定的功能/动作的各方面的指令。

[0130] 计算机可读程序指令还可以被加载到计算机、其他可编程数据处理装置或其他设备上，以使得在计算机、其他可编程装置或其他设备上执行一系列操作步骤，以产生计算机实现的过程，使得在计算机、其他可编程装置或其他设备上执行的指令实现流程图和/或框图的一个或多个框中指定的功能/动作。

[0131] 附图中的流程图和框图示出了根据本发明的各种实施例的系统、方法和计算机程序产品的可能实现的架构、功能和操作。在这点上，流程图或框图中的每个框可以表示指令的模块、段或部分，其包括用于实现指定的逻辑功能的一个或多个可执行指令。在一些替代实施方案中，框中所注明的功能可不按图中所注明的次序发生。例如，连续示出的两个框实际上可以作为一个步骤来实现，同时、基本同时、以部分或全部时间重叠的方式执行，或者这些框有时可以以相反的顺序执行，这取决于所涉及的功能。还将注意，框图和/或流程图图示的每个框以及框图和/或流程图图示中的框的组合可以由执行指定功能或动作或执行专用硬件和计算机指令的组合的专用的基于硬件的系统来实现。

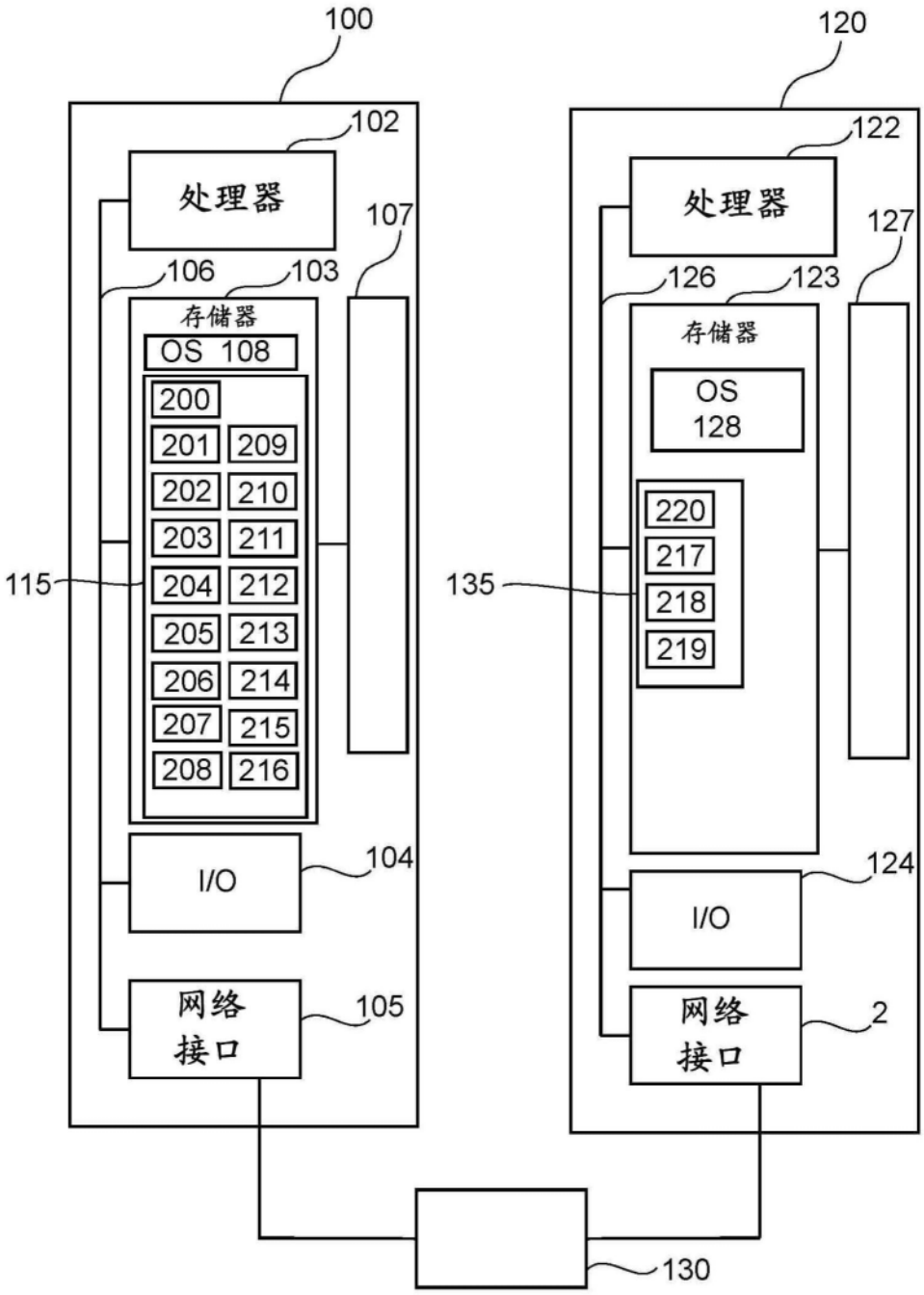

100
120
102
122
处理器
处理器
107
127
106
103
126
123
存储器
存储器
OS 108
OS
128
200
201
209
202
210
203
211
220
217
218
219
115
204
212
135
205
213
206
214
207
215
208
216
I/O
104
I/O
124
网络
接口
105
网络
接口
2
130

图1

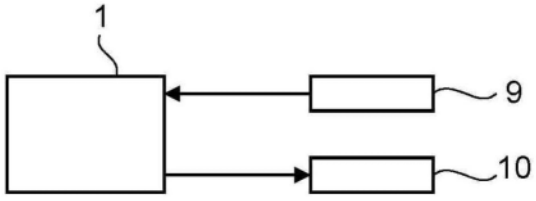

1
9
10

图2

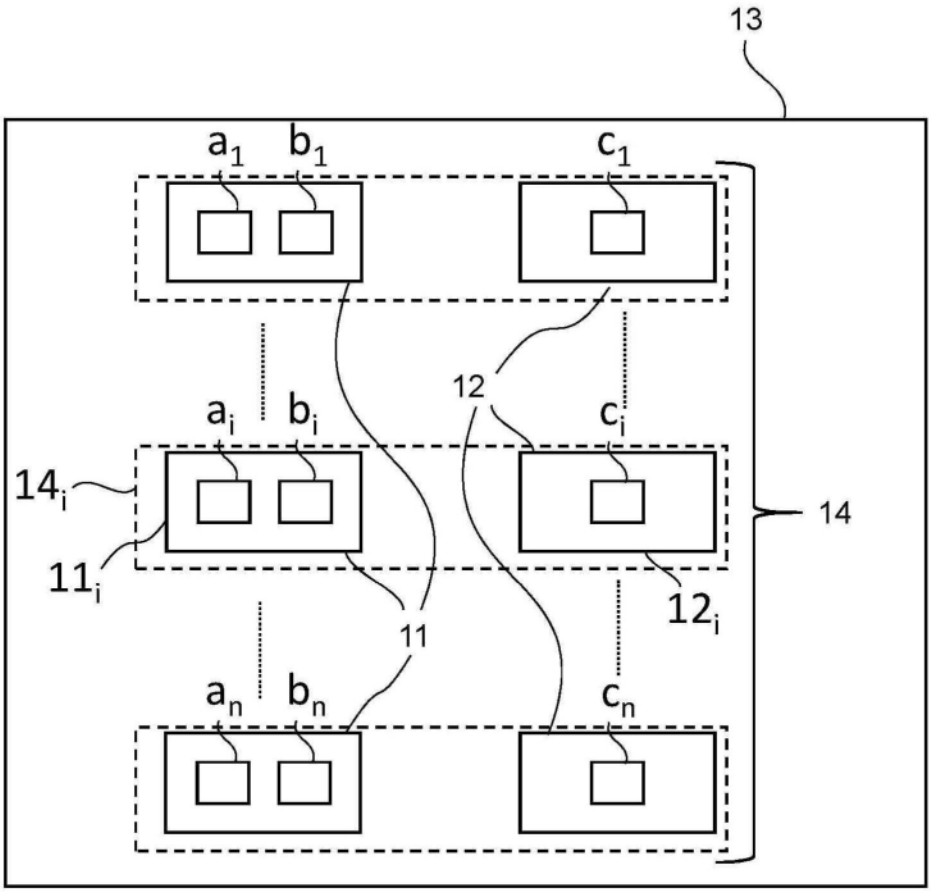


图3

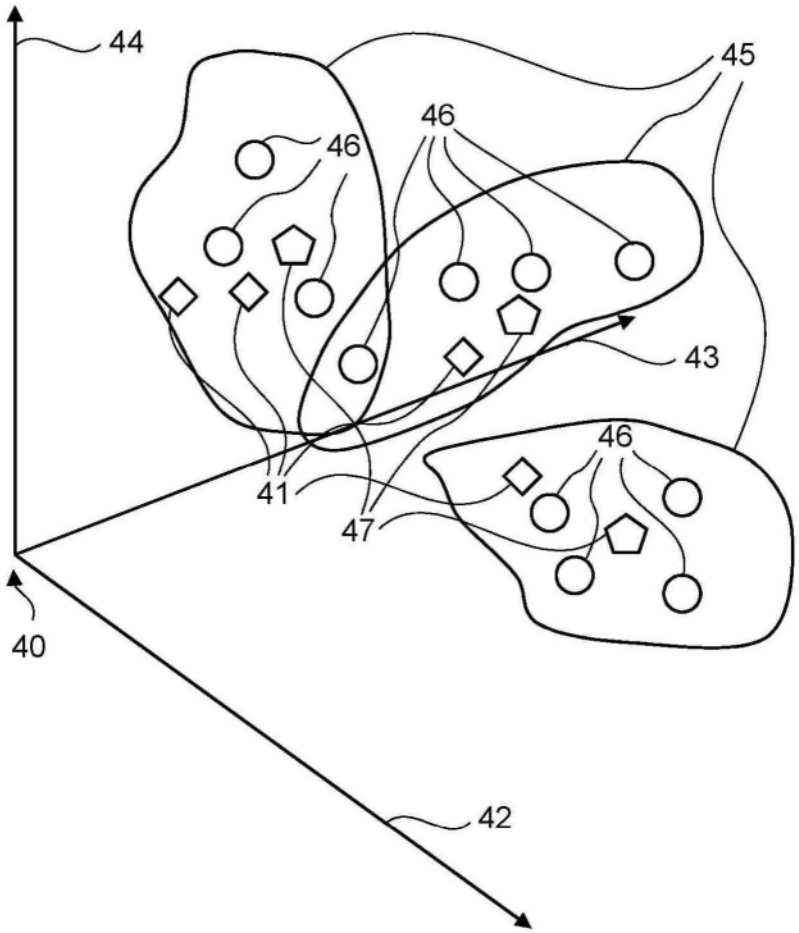


图4

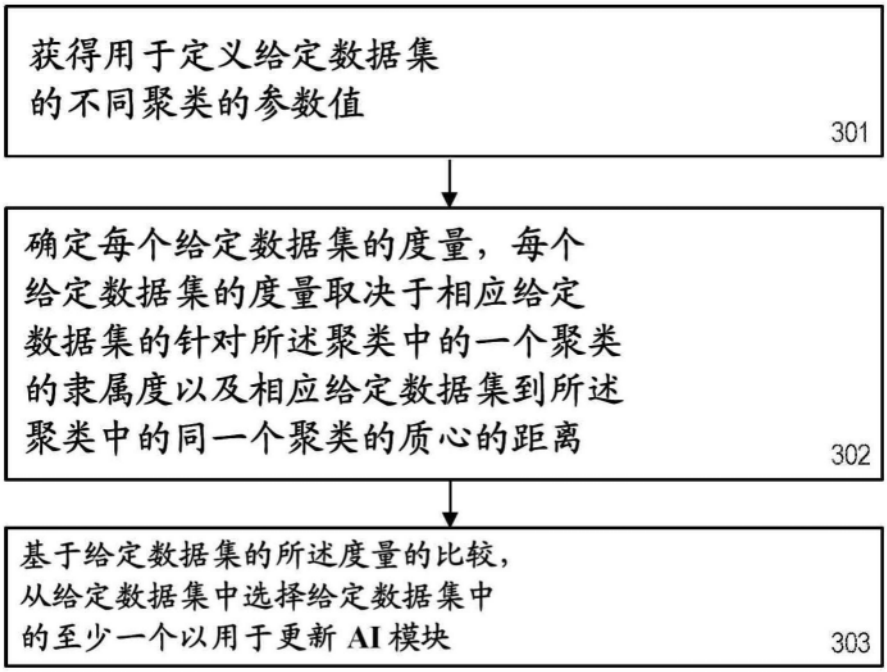

获得用于定义给定数据集
的不同聚类的参数值
301
确定每个给定数据集的度量，每个
给定数据集的度量取决于相应给定
数据集的针对所述聚类中的一个聚类
的隶属度以及相应给定数据集到所述
聚类中的同一个聚类的质心的距离
302
基于给定数据集的所述度量的比较，
从给定数据集中选择给定数据集中
的至少一个以用于更新 AI 模块
303

图5