(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112347322 B
(45)授权公告日 2024.11.05

(21)申请号 202010783314.9

(22)申请日 2020.08.06

(65)同一申请的已公布的文献号
申请公布号 CN 112347322 A

(43)申请公布日 2021.02.09

(30)优先权数据
16/532,505 2019.08.06 US

(73)专利权人 国际商业机器公司
地址 美国纽约

(72)发明人 G·埃佐夫 A·法尔卡什
A·高登斯特恩 R·施梅尔金
M·G·莫夫尔

(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 郑宗玉

(51)Int.Cl.
G06F 16/906(2019.01)
G06F 18/214(2023.01)
G06N 3/0464(2023.01)

(56)对比文件
CN 107085585 A,2017.08.22
CN 108604902 A,2018.09.28

审查员 张曼

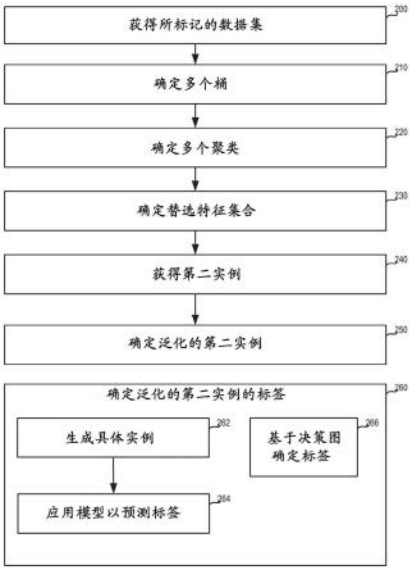
权利要求书2页 说明书12页 附图4页

(54)发明名称

用于预测模型的数据泛化

(57)摘要

用于预测模型的数据泛化。用于预测模型的数据泛化的方法、装置和产品。该方法包括:基于所标记的数据集,确定多个桶,多个桶中的每个桶具有相关联的标签;确定多个聚类,将相同桶中的相似实例进行分组;基于多个聚类,确定包括泛化特征集合的替选特征集合,其中每个泛化特征对应于多个聚类中的聚类,其中对应于聚类的泛化特征指示被映射到对应聚类的实例;获得第二实例;确定泛化的第二实例,泛化的第二实例包括对第二实例的替选特征集合的评估;以及基于泛化的第二实例,确定第二实例的标签。



CN 112347322 B

1.一种方法,包括:

获得所标记的数据集,其中所标记的数据集与年龄相关,所标记的数据集包括多个实例及其标签,所述标签包括年龄标签,其中每个实例包括对特征集合的评估;

基于标签,确定多个桶,所述多个桶中的每个桶具有相关联的标签,其中每个桶将来自所标记的数据集中的具有在与桶的相关联的标签的相似性阈值之内的标签的实例进行分组;

确定多个聚类,其中每个聚类包括由相同桶包括的多个实例,其中所述确定多个聚类是基于对实例的特征集合的评估,由此将相似的实例分组到聚类中;

基于所述多个聚类,确定包括泛化特征集合的替选特征集合,其中所述泛化特征集合中的每个泛化特征对应于所述多个聚类中的聚类,其中对应于聚类的泛化特征指示作为对应聚类的成员的实例;

获得第二实例;

确定泛化的第二实例,其中泛化的第二实例包括对第二实例的替选特征集合的评估;以及

基于泛化的第二实例,确定第二实例的标签。

2.根据权利要求1所述的方法,其中所述相似性阈值是相等性阈值,由此每个桶将来自所标记的数据集中的具有每个桶的相关联的标签的实例进行分组。

3.根据权利要求1所述的方法,其中所述确定第二实例的标签包括:

基于对泛化的第二实例的替选特征集合的评估,确定所识别的聚类,其中第二实例被映射到所识别的聚类;以及

其中第二实例的标签是基于与所识别的聚类相关联的桶的相关联的标签来确定的。

4.根据权利要求3所述的方法,其中所述确定所识别的聚类包括确定多个所识别的聚类,其中第二实例的标签是基于与所述多个所识别的桶相关联的一个或多个桶的一个或多个标签来确定的。

5.根据权利要求1所述的方法,其中所述确定第二实例的标签包括:

基于泛化的第二实例确定具体的第二实例;以及

将具体的第二实例提供给预测模型,其中所述预测模型被配置为提供具体的第二实例的第二标签。

6.根据权利要求5所述的方法,其中具体的第二实例不同于第二实例。

7.根据权利要求5所述的方法,其中具体的第二实例是由聚类包括的所观测的实例的统计表示。

8.根据权利要求5所述的方法,其中聚类包括质心,其中质心在聚类内部,其中具体的第二实例是聚类的质心。

9.根据权利要求1所述的方法,

其中获得第二实例和确定泛化的第二实例是在边缘设备上执行的;

其中所述方法还包括:由所述边缘设备将泛化的第二实例传送到服务器;以及

其中基于泛化的第二实例确定第二实例的标签是在服务器上执行的。

10.根据权利要求1所述的方法,其中所述获得所标记的数据集包括:

获得多个实例;以及

将预测模型应用于所述多个实例中的每个实例以确定其标签,其中所述预测模型被配置为确定实例的标签。

11.一种保存程序指令的非暂时性计算机可读存储介质,所述程序指令在被处理器读取时使所述处理器执行根据权利要求1-10中的任一项所述的方法。

12.一种具有处理器和耦合的存储器的计算机化的装置,所述处理器适于执行根据权利要求1-10中的任一项所述的方法的步骤。

13.一种系统,所述系统包括分别用于执行根据权利要求1-10中的任一项所述的方法的步骤的模块。

用于预测模型的数据泛化

技术领域

[0001] 本公开总体上涉及对数据进行泛化,并且尤其涉及使用聚类对用于预测模型的数据进行泛化。

背景技术

[0002] 数据最小化可以是指将个人信息的收集限制为与实现特定目的直接相关并且为实现特定目的所需要的信息的实践。随着公司和组织开始理解数据的力量,并且随着数据变得更加无处不在并且更易于收集,分析师面临大量的数据。有一段时间,人们的冲动是无限期地保存所有这些数据。随着智能电话、物联网(IoT)设备等的快速采用,组织面临越来越多的方式来收集越来越多种类的数据,包括尤其是私人的、个人可识别的数据。数据管理员现在不再保存所有数据,而是采用数据最小化策略,只保留相关和需要的数据。

发明内容

[0003] 所公开的主题的一个示例性实施例是一种方法,包括:获得所标记的数据集,其中所标记的数据集包括多个实例及其标签,其中每个实例包括对特征集合的评估;基于标签,确定多个桶,所述多个桶中的每个桶具有相关联的标签,其中每个桶将来自所标记的数据集中的具有在与桶的相关联的标签的相似性阈值之内的标签的实例进行分组;确定多个聚类,其中每个聚类包括由相同桶包括的多个实例,其中所述确定多个聚类是基于对实例的特征集合的评估,由此将相似的实例分组到聚类中;基于所述多个聚类,确定包括泛化特征集合的替选特征集合,其中所述泛化特征集合中的每个泛化特征对应于所述多个聚类中的聚类,其中对应于聚类的泛化特征指示作为对应聚类的成员的实例;获得第二实例;确定泛化的第二实例,其中泛化的第二实例包括对第二实例的替选特征集合的评估;以及基于泛化的第二实例,确定第二实例的标签。

[0004] 所公开的主题的另一示例性实施例是一种计算机程序产品,包括保存程序指令的非暂时性计算机可读存储介质,所述程序指令在被处理器读取时使所述处理器执行:获得所标记的数据集,其中所标记的数据集包括多个实例及其标签,其中每个实例包括对特征集合的评估;基于标签,确定多个桶,所述多个桶中的每个桶具有相关联的标签,其中每个桶将来自所标记的数据集中的具有在与桶的相关联的标签的相似性阈值之内的标签的实例进行分组;确定多个聚类,其中每个聚类包括由相同桶包括的多个实例,其中所述确定多个聚类是基于对实例的特征集合的评估,由此将相似的实例分组到聚类中;基于所述多个聚类,确定包括泛化特征集合的替选特征集合,其中所述泛化特征集合中的每个泛化特征对应于所述多个聚类中的聚类,其中对应于聚类的泛化特征指示作为对应聚类的成员的实例;获得第二实例;确定泛化的第二实例,其中泛化的第二实例包括对第二实例的替选特征集合的评估;以及基于泛化的第二实例,确定第二实例的标签。

[0005] 所公开的主题的又一示例性实施例是一种具有处理器和耦合的存储器的计算机化的装置,所述处理器适于执行以下步骤:获得所标记的数据集,其中所标记的数据集包括

多个实例及其标签,其中每个实例包括对特征集合的评估;基于标签,确定多个桶,所述多个桶中的每个桶具有相关联的标签,其中每个桶将来自所标记的数据集中的具有在与桶的相关联的标签的相似性阈值之内的标签的实例进行分组;确定多个聚类,其中每个聚类包括由相同桶包括的多个实例,其中所述确定多个聚类是基于对实例的特征集合的评估,由此将相似的实例分组到聚类中;基于所述多个聚类,确定包括泛化特征集合的替选特征集合,其中所述泛化特征集合中的每个泛化特征对应于所述多个聚类中的聚类,其中对应于聚类的泛化特征指示作为对应聚类的成员的实例;获得第二实例;确定泛化的第二实例,其中泛化的第二实例包括对第二实例的替选特征集合的评估;以及基于泛化的第二实例,确定第二实例的标签。

附图说明

[0006] 通过结合附图进行的以下详细描述,将更充分地理解和领会本公开的主题,在附图中,对应或相似的数字或字符指示对应或相似的组件。除非另外指出,否则附图提供本公开的示例性实施例或方面,而不限制本公开的范围。在附图中:

[0007] 图1A和图1B示出根据所公开的主题的一些示例性实施例的桶和聚类;

[0008] 图1C示出根据所公开的主题的一些示例性实施例的决策图;

[0009] 图1D示出根据所公开的主题的一些示例性实施例的聚类的图示;

[0010] 图2示出根据所公开的主题的一些示例性实施例的方法的流程图;以及

[0011] 图3示出根据所公开的主题的一些示例性实施例的装置的框图。

具体实施方式

[0012] 由所公开的主题处理的一个技术问题是最小化从用户收集的数据。在一些示例性实施例中,收集数据的实体可以服从通用数据保护法规(GDPR)法规。因此,可能要求实体将数据收集限制为与数据可被处理的目的相关的需要的数据。附加地或替选地,GDPR和类似法规可能要求以某些方式存储和保护所收集的私人数据。因此,减少所存储的数据量可以减少实体的责任。

[0013] 由所公开的主题处理的另一技术问题是最小化正在针对决策模型收集的数据。在一些情况下,可能期望最小化数据,例如正在收集的特征的数量及其各自的粒度,同时仍然能够利用决策模型来提供质量决定、预测等。作为示例,相同的数据记录可以由被配置为预测心脏病发作的预测模型和被配置为预测人的后代数量的预测模型利用。可以对于每个决策模型应用不同类型的数据最小化,例如不同类型的特征泛化。作为示例,在年龄特征被泛化到5年的范围的情况下,第一预测模型可以具有90%以上的性能度量,而在年龄特征被泛化到这样的范围的情况下,第二预测模型可以具有80%以下的性能度量。作为另一个示例,一个模型可能对50岁以上的年龄及其精确值敏感,而另一个模型可能无动于衷并且为40岁以上的所有值提供相似的预测。在一些情况下,决策模型可以是手动模型,该手动模型被实体用于做出其决定,诸如例如,由银行家做出的批准或拒绝贷款的决定,由银行家做出的关于授权最大信用额的决定,由办事员做出的提供对要供应的服务的报价的决定等。

[0014] 由所公开的主题处理的又一技术问题是在不影响预测模型的性能度量的情况下最小化提供给预测模型的数据。可能期望在不使预测模型的性能度量降低到阈值以下的情

况下最小化数据。阈值可以是绝对阈值,例如90％,92％等。附加地或替选地,阈值可以是相对于在数据最小化之前的预测模型的性能度量的相对阈值。作为示例,可能期望在最小化数据之后,性能度量的下降不得超过5％。

[0015] 一个技术方案是确定由决策模型利用的特征集合的泛化。在一些示例性实施例中,可以不是确定该特征集合,而是可以确定替选特征集合。替选特征集合可以包括泛化特征集合。泛化特征可以是特征集合中的一个或多个特征的泛化。对特征集合进行泛化可以使得数据最小化。

[0016] 在一些示例性实施例中,可以使用聚类处理来确定替选特征集合。泛化特征可以与在聚类处理期间确定的聚类相关联。

[0017] 另一技术方案是执行两阶段聚类处理。第一阶段可以基于由所标记的数据集包括的标签,并且可以产生多个桶,所述多个桶中的每个桶可以与标签的值相关联。第二阶段可以基于对由所标记的数据集包括的实例的特征的评估。

[0018] 在一些示例性实施例中,第二阶段可以应用于由至少一个桶包括的实例。可以利用诸如K均值、分层聚类、仿射传播(Affinity Propagation)等聚类算法。

[0019] 在一些示例性实施例中,桶可以将来自所标记的数据集中的具有在与该桶相关联的标签的相似性阈值之内的标签的实例进行分组。作为示例,桶可以与30岁的年龄标签相关联。相似性阈值可以定义2年的范围,从而使得可以将具有在29岁到31岁之间的年龄标签的实例分组到桶。在一些示例性实施例中,相似性阈值可以是相等性阈值、同一性阈值等,从而使得桶将具有完全相同的标签的实例进行分组。

[0020] 注意,将实例分组到桶可以基于标签的相似性阈值,而确定聚类可以基于对特征的评估的相似性在每个桶处执行。附加地或替选地,聚类可以将由相同桶包括的实例进行分组。在一些示例性实施例中,每个聚类可以与桶相关联。附加地或替选地,对于单个桶,可能存在多个所确定的聚类。

[0021] 在一些示例性实施例中,可以获得第二实例,并且可以确定第二实例的标签。第二实例可以是看不见的实例,例如不是由所标记的数据集包括的实例,标签不可用并且需要确定标签的实例等。在一些示例性实施例中,可以在没有已知的对应标签的情况下获得第二实例。在一些情况下,对第二实例进行泛化可以基于对第二实例的替选特征集合的评估。对替选特征集合的评估可以产生第二实例的对应的泛化实例。泛化实例可以指示包括第二实例的一个或多个聚类,也被称为所识别的聚类。作为示例,第二实例可以被映射到单个聚类(也被称为单个聚类的“成员”)。附加地或替选地,第二实例可以被映射到一个以上的聚类。在一些示例性实施例中,可以基于一个或多个所识别的聚类的标签来确定泛化实例的标签。作为示例,如果泛化实例指示映射到单个聚类,则可以将与其相关联的标签、与其相关联的标签的平均值等确定为泛化实例的标签。作为另一示例,如果泛化实例指示在两个或更多个聚类中的成员资格,则可以基于所识别的聚类的至少一部分的一个或更多个标签来确定标签。在一些示例性实施例中,成员资格度量可以被计算并且被用于确定标签。成员资格度量可以指示第二实例与哪个聚类最关联。作为示例,可能期望将代表人的实例标记为律师或专利代理师。该实例可以具有指示收益渠道的至少两个特征,例如指示人的来自诉讼的收入的特征和指示人的来自起草专利申请的收入的特征。可以确定几个泛化特征,每个泛化特征可以对应于不同的聚类。一些聚类可以包括被标记为“律师”的实例,并且一

些聚类可以包括被标记为“专利代理师”的实例。要标记的实例可以被分析并且被确定为映射到两个不同的聚类,第一聚类与“律师”标签相关联,而第二聚类与“专利代理师”标签相关联。可以定义成员资格度量,以测量每个聚类中的实例的成员资格。作为示例,可以通过测量每个收益渠道来定义成员资格函数。在人的来自诉讼的收入高于该人的来自起草的收入(例如,具有或没有归一化)的情况下,可以确定该实例更多是律师聚类中的成员,并且该实例可以被标记为律师。另一方面,在人的来自起草的收入高于该人的来自诉讼的收入的情况下,可以确定,与律师聚类相比,该实例更多是专利代理师聚类中的成员,并且该实例可以被标记为专利代理师。可能会注意到,成员资格度量可能会在两个聚类上返回相同的度量(例如,人的来自诉讼的收入等于该人的来自起草的收入,或者两个归一化收入相等)。在该情况下,实例可能是离群值,并且可以在两个选项之间随机确定标签。

[0022] 附加地或替代地,通过测量与每个聚类的质心的距离,可以确定该人是更多代表专利代理师还是律师。在一些情况下,泛化特征可以包括与每个聚类的质心的距离度量。可以基于实例最接近的聚类(例如,对应的泛化特征的值最低的所识别的聚类)的标签来确定标签。附加地或替选地,可以基于加权平均值或其他计算来确定标签,所述其他计算考虑到实例与每个所识别的聚类的相似性度量,如可能由泛化特征的值所展示的。

[0023] 在一些示例性实施例中,为了确定实例的标签,可以确定具体实例。可以将实例泛化为泛化实例,以便最小化数据暴露。基于泛化实例,可以确定具体实例。具体实例可以包括对原始特征的评估,而不是对用于代表泛化的替选特征集合的评估。可以将具体实例提供给决策模型,以便确定该实例的标签。决策模型可以是自动模型、预测模型、由人执行的手动决定处理等。在一些情况下,具体实例可以是所观测的实例的统计表示,所观测的实例是所识别的聚类的成员。所观测的实例可以是在实践所公开的主题期间先前观测到的实例,例如由所标记的数据集包括,作为训练数据集的一部分,提供给决策模型等。所观测的实例的统计表示可以是所观测的实例的平均值、所观测的实例的模式、所观测的实例的均值等。附加地或替选地,包括实例的聚类可以具有质心,并且聚类可以包括质心。包括其质心的聚类的示例可以是凸聚类。在该情况下,质心可能是具体值。附加地或替选地,在实例被多个所识别的聚类包括的情况下,可以执行基于所观测的实例、他们中的每个的质心等的计算以生成具体实例。作为另一示例,所生成的具体实例可以被确定为被包括原始实例的所有聚类包括,诸如可以基于所有所识别的聚类的交集来生成。可以基于在感兴趣的聚类中观测到的实例等,将具体实例生成为相交的聚类的质心。

[0024] 在一些示例性实施例中,在所观测的实例的数量低于阈值的情况下,具体实例可以是质心。一旦所观测的实例的数量超过阈值,则具体实例可以是所观测的值的统计表示。

[0025] 在一些示例性实施例中,可以使用预测模型来确定由所公开的主题利用的所标记的数据集。在一些示例性实施例中,可以获得未标记的数据集。预测模型可以应用于未标记的数据集。使用预测模型,可以确定未标记的数据集中的每个实例的预测标签,从而产生所标记的数据集。这可能与常规的所标记的数据集相反,常规的所标记的数据集通常利用准确并且正确的标记而不是利用预测的标记来标记。

[0026] 在一些示例性实施例中,可以获得实例并且在边缘设备上对该实例进行泛化。实例的泛化可以被传输到服务器,并且可以在服务器上确定第二实例的标签。

[0027] 利用所公开的主题的一个技术效果是,决策模型为了确定标签而可能需要的数据

被最小化。代替泄露完整的信息,可以泄露其最小化的表示,从而减少了保存和处理的私有信息。

[0028] 利用所公开的主题的另一技术效果是,减少了为预测实例的标签所需要的计算资源。在一些示例性实施例中,代替地,如果应用预测模型,则可以基于所识别的聚类来应用相对非能力密集的计算。在一些情况下,预测模型可以是ANN(人工神经网络)模型(例如CNN(卷积神经网络)、RNN(递归神经网络)、DNN(深度神经网络)等)、非ANN模型(例如决策树、SVM(支持向量机)等)。在一些示例性实施例中,与应用预测模型本身所需要的资源相比,所公开的主题可以使用减少量的计算资源来提供预测。

[0029] 利用所公开的主题的另一技术效果是,减少了存储数据所需要的存储空间、传输数据所需要的带宽以及传输数据所需要的功率。在一些示例性实施例中,可以保存数据实例以用于未来的使用,例如特定于域的使用、质量保证、其他模型的训练、重新训练预测模型等。附加地或替选地,可能期望将数据实例传输到远程服务器。作为示例,可以从IoT设备获得数据实例。IoT设备可以每秒测量一次温度、湿度、光线等。代替保存精确值,可以应用重新编码以利用减少量的比特来表示更泛化的数据。在一些情况下,泛化可以被视为有损压缩的一种形式,它丢失了与由从IoT设备收集数据的服务器所使用的决策模型有关的无关紧要的信息。

[0030] 所公开的主题可以提供对任何现有技术和本领域中以前已成为常规或传统的任何技术的一种或多种技术改进。

[0031] 鉴于本公开,附加技术问题、解决方案和效果对于本领域普通技术人员而言可能是明显的。

[0032] 现在参考图1A和图1B,图1A和图1B示出根据所公开的主题的一些示例性实施例的桶和聚类。

[0033] 图1A示出了实例,例如实例103。每个实例包括特征集合中的每个特征的值。例如,实例103具有评估{年龄=30,性别=M,头发颜色=棕色}。每个实例可以与标签相关联,例如关于布尔属性的真或假、工资属性等的标签。

[0034] 可以将实例分配到桶中,例如桶101、102。每个桶包括具有相同标签或具有相似标签的实例,例如具有不超过桶的标签的阈值的相似性度量的标签。例如,桶101可以包括具有真标签的所有实例,而桶102可以包括具有“假”标签的所有实例。作为另一示例,桶101可以包括具有30K-40K的工资的所有实例,而桶102可以包括具有40K-50K以上的工资的所有实例。

[0035] 图1B示出了将由桶101包括的实例聚类为三个聚类:聚类111、112和113。聚类可以将具有与特征集合相似的评估的实例进行分组。例如,聚类111可以包括具有相似的头发颜色和相似的在30岁至35岁之间的年龄的实例。在一些示例性实施例中,可以不要求年龄在30岁至35岁之间的具有棕色或黑色头发的人公开其确切年龄或头发颜色,同时仍然允许所公开的主题提供其相对精确的标签。由于它们是桶101中的聚类的成员,因此可以基于桶101的标签来确定标签,而不需要与年龄和头发颜色有关的精确信息。附加地或替选地,可以生成类似地位于聚类111内的潜在替选的具体实例,并且例如可以通过对潜在替选的具体实例应用决策模型来使用这样的具体实例来确定实际实例的标签。注意,在图1B的所示的示例中,可能不会基于当前数据、当前聚类等对其他年龄范围进行泛化。作为示例,年龄

在40岁至60岁之间的人可能必须透露其确切年龄。

[0036] 在一些示例性实施例中,数据集越大并且数据中存在的离群值越少,可以使用聚类进行泛化更有可能。

[0037] 在一些示例性实施例中,可以通过指示实例属于哪个聚类来提供泛化。附加地或替选地,泛化可以包括对于每个聚类的对该实例是否属于该聚类的指示、以及与之有关的成员资格度量。

[0038] 在一些示例性实施例中,所生成的聚类(以及因此替选特征集合)可能没有覆盖整个域,因此有可能稍后可以接收没有映射到任何聚类中的数据点。在该情况下,有可能:或者不对这样的数据点进行泛化,或者将其映射到最接近的聚类,因此实现泛化但是可能会失去一些准确性。

[0039] 现在参考图1C,图1C示出根据所公开的主题的一些示例性实施例的决策图的图示。

[0040] 可以基于桶和聚类(例如,图1A和图1B的桶和聚类)来确定决策图120。决策图120可以示出与图1A和图1B中的特征集合相对应的替选特征集合。图1A和图1B示出了具有3个特征的示例。为了简单起见,决策图120示出了2个替选特征:年龄和性别。决策图120对年龄特征进行泛化。年龄特征使用以下四个子域来表示:(20-25)、(25-30)、(30-35)、(35-40)。每个区域可以对应于不同的标签。区域122或区域132中的实例可以具有第一标签,区域124、区域128或区域134中的实例可以具有第二标签,而区域126中的实例可以具有第三标签。作为示例,实例可能具有对于性别特征的“女性”评估、以及作为对于年龄特征的评估的23岁的值。该实例可以位于区域122并且具有第一标签。

[0041] 在一些示例性实施例中,对于一些特征,可以进一步对其他特征进行泛化,从而产生条件泛化。作为示例,如可以在决策图120中看到的,在实例具有对于性别的“女性”值的情况下,泛化的年龄特征可以具有以下两个子域而不是四个子域:(20-30)和(30-40)。在一些示例性实施例中,可以使用所公开的主题以便最小化从用户获得的数据。在用户是女性的情况下,可能足以知道她的年龄是在20岁到30岁之间还是在30岁到40岁之间。

[0042] 在一些示例性实施例中,可以基于聚类来生成决策图120。在一些示例性实施例中,决策图120可以包括区域,每个区域对应于不同的聚类。在这样的情况下,区域可能具有潜在的不规则形状,可能重叠,空间的部分可能不与任何区域相关联等。在一些示例性实施例中,可以例如通过定义分隔与不同标签相关联的聚类的线来生成决策图120作为聚类的抽象。在一些示例性实施例中,可以利用直线将由决策图120定义的空间划分为区域,使得每个区域基本上包括与相同标签相关联的聚类。在一些情况下,可以利用与轴平行的线,以便提供将轴的值明确地划分为子域。

[0043] 现在参考图1D,图1D示出根据所公开的主题的一些示例性实施例的聚类的图示。

[0044] 在一些示例性实施例中,可以示出二维特征空间以可视地示出聚类。聚类可以是特征空间中的形状。在二维空间中,该形状可以是二维形状,该二维形状可以是对称的或者可以是不对称的。聚类152以圆形的形式例示聚类。作为聚类152的中心的质心152c在聚类152内部。当计算由聚类152包括的泛化实例的标签时,可以使用质心152c作为具体实例。附加地或替选地,聚类156可以不是凸的分组。聚类156的质心156c在聚类156的外部。在一些示例性实施例中,当生成聚类156的具体实例时,质心156c可以不按原样地使用。在一些情

况下,可以选择替选的所观测的实例。附加地或替选地,可以使用最接近于质心156c的所观测的实例。附加地或替选地,仅在观测到聚类的质心时才可以使用它。否则,可以使用在聚类内并且最接近于聚类的替选的所观测的实例。

[0045] 现在参考图2,图2示出根据所公开的主题的一些示例性实施例的方法的流程图。

[0046] 在步骤200,可以获得所标记的数据集。所标记的数据集可以包括实例及其标记。在一些示例性实施例中,可以通过将决策模型应用于未标记的实例的数据集来获得所标记的数据集的标签。在一些示例性实施例中,决策模型可以是被配置为确定实例的标签的预测模型,从而提供可能准确或可能不准确的预测标签。附加地或替选地,所标记的数据集可以是从任何其他来源或使用任何其他方法收集的所标记的数据集。

[0047] 在步骤210,可以确定多个桶。每个桶可以与标签关联。在一些示例性实施例中,每个桶可以将由具有相同标签的所标记的数据集包括的实例进行分组。附加地或替选地,桶可以包括具有相似但是不一定相同的标签的实例。在一些示例性实施例中,实例的分组可以基于相似性阈值。阈值可以定义实例的标签和桶的相关联的标签应该有多相似。在一些示例性实施例中,相似性阈值可以是相等性阈值,从而使得只有具有与桶相关联的标签确切地相同的实例才可以被分组到桶。在一些示例性实施例中,可以基于与桶相关联的标签并且基于每个实例的标签来计算相似性度量。可以将相似性度量与相似性阈值进行比较,以确定实例是否为桶的成员。

[0048] 在步骤220,可以确定多个聚类。可以对于每个桶确定多个聚类。在一些示例性实施例中,可以对于每个桶分别地并且独立地确定多个聚类。在一些示例性实施例中,单个实例可以是一个以上聚类的成员。附加地或替选地,这些聚类之间可以具有或可以不具有重叠。在一些示例性实施例中,每个聚类可以恰好被一个桶包括。每个聚类可以包括相似的实例。关于分组到聚类中的实例的相似性可以基于对每个实例的特征集合的评估。在一些示例性实施例中,可以使用任何聚类技术来执行聚类的确定,例如但不限于,k均值聚类、基于质心的聚类、基于分布的聚类、具有噪声的应用的基于密度的空间聚类(DBSCAN)、使用分层的平衡迭代减少和聚类(BIRCH)等。

[0049] 在步骤230,可以确定替选特征集合。在一些示例性实施例中,替选特征集合可以是特征集合的泛化。在一些示例性实施例中,替选特征集合可以包括特征的一部分、一个或多个泛化的特征等。附加地或替选地,替选特征集合可以包括泛化特征。泛化特征可以对应于由特征集合包括的特征。特征可以具有可能值的范围。泛化特征可以具有缩小的域,包括更少数量的可能值,每个可能值对应于该域中的子域。在一些示例性实施例中,可以基于诸如在步骤220确定的聚类之类的聚类来确定泛化特征。在一些示例性实施例中,泛化特征可以是指示实例对聚类的成员资格的特征。在一些示例性实施例中,泛化特征可以指示成员资格和成员资格度量。例如,代替包括50个不同特征的特征集合,替选集合可以包括10个替选特征,每个替选特征指示实例对不同聚类的成员资格。可以注意到,在一些情况下,可能存在比原始特征更多的替选特征。在一些示例性实施例中,使用替选特征集合可以提供用户的私有信息的减少的泄露。即使替选特征的数量大于原始特征的数量,这样的减少的泄露也可以是适用的。

[0050] 在步骤240,可以获得第二实例。第二实例可以是未标记的实例。在一些示例性实施例中,可能期望确定第二实例的标签。在一些示例性实施例中,第二实例可以是先前没有

被所公开的主题使用的实例,诸如没有被包括在步骤200的所标记的数据集中。附加地或替选地,第二实例可以是没有被用来训练根据所公开的主题而使用的预测模型的实例。在一些示例性实施例中,第二实例可以例如基于在第二实例中定义的对特征集合的评估而被映射到一个或多个所识别的聚类。

[0051] 在步骤250,可以确定泛化的第二实例。泛化的第二实例可以是第二实例的泛化。在一些示例性实施例中,泛化的第二实例可以包括对替选特征集合的评估。在一些示例性实施例中,可以通过将对第二实例的特征的评估映射到对替选特征集合的评估来确定泛化的第二实例。附加地或替选地,可以基于在步骤220确定的第二实例对不同聚类的成员资格来确定泛化的第二实例。

[0052] 在步骤260,可以确定泛化的第二实例的标签。为泛化的第二实例确定的标签可以用作第二实例的标签。在一些示例性实施例中,确定泛化的第二实例的标签可以包括执行步骤262和步骤264。附加地或替选地,确定泛化的第二实例的标签可以包括执行步骤266。

[0053] 在步骤262,可以生成具体实例。可以基于泛化的第二实例来生成具体实例。具体实例可以包括对包括与泛化的第二实例一致的特征集合的特征的评估。在一些示例性实施例中,泛化的第二实例可以例如经由替选特征集合的值来指示包括第二实例的一个或多个聚类。可以生成具体实例,以便提供作为包括第二实例的一个或多个聚类的成员的实例。具体实例可能潜在地与第二实例不同。在一些示例性实施例中,对于具体实例和第二实例两者,可以利用相同的泛化--泛化的第二实例。

[0054] 在一些示例性实施例中,具体实例可以是包括第二实例的聚类的质心。附加地或替选地,可以在验证质心也是聚类的成员之后利用质心。附加地或替选地,为了避免利用虚假的和不现实的值,可以利用所观测的实例。所观测的实例可以是在公开的主题的应用期间观测到的实例,例如在所标记的数据集中的实例、所获得的要被标记的实例(例如,步骤240的第二实例)等。在一些示例性实施例中,所观测的实例可以被选择并且被用作具体实例。该选择可以基于所观测的实例与质心之间的关系或所有所观测的实例的另一种统计表示。例如,可以选择和利用最接近于质心并且被包括在聚类内的所观测的实例。附加地或替选地,在泛化的第二实例指示对多于单个聚类的成员资格的情况下,可以从作为与第二实例相同的聚类集合的成员的所观测的实例中选择具体实例。附加地或替选地,可以基于成员资格度量来确定顶部聚类,并且可以生成顶部聚类中的具体实例。

[0055] 在步骤264,可以将模型应用于具体实例,以便预测第二实例的标签。在一些示例性实施例中,该模型可以是用于对步骤200的所标记的数据集进行标记的决策模型。附加地或替选地,该决策模型可以是例如使用机器学习技术实现的自动化模型、手动模型类似。

[0056] 在步骤266,可以基于诸如决策图120之类的决策图来确定标签。决策图中的每个区域可以对应于标签。通过确定对替选特征集合的评估,可以将泛化实例映射到区域,并且可以基于该区域来确定标签。

[0057] 附加地或替选地,可以基于包括第二实例的聚类的标签来确定标签。附加地或替选地,在存在包括第二实例的多个聚类的情况下,可以例如通过计算其加权平均值来利用其标签。加权平均值的权重可以基于如可以记录在泛化的第二实例中的第二实例对不同聚类的成员资格度量。

[0058] 现在参考图3,图3示出根据所公开的主题的一些示例性实施例的装置的框图。

[0059] 在一些示例性实施例中,装置300可以包括一个或多个处理器302。处理器302可以是中央处理单元(CPU)、微处理器、电子电路、集成电路(IC)等。处理器302可用来执行由装置300或其任何子组件所需要的计算。

[0060] 在所公开的主题的一些示例性实施例中,装置300可以包括输入/输出(I/O)模块305。I/O模块305可以用来向用户提供输出并且从用户接收输入,诸如例如,获得泛化实例、提供预测等。在一些示例性实施例中,I/O模块305可以被配置为获得预测模型、获得数据集、获得所标记的数据集等。附加地或替选地,I/O模块305可以被配置为发送泛化实例。

[0061] 在一些示例性实施例中,装置300可以包括存储器单元307。存储器单元307可以是硬盘驱动器、闪存盘、随机存取存储器(RAM)、存储器芯片等。在一些示例性实施例中,存储器单元307可以保存可操作以使处理器302执行与装置300的任何子组件相关联的动作的程序代码。在一些示例性实施例中,存储器单元307可以存储数据集、度量性能结果等。附加地或替选地,存储器单元307可以存储预测模型、桶、聚类等。

[0062] 存储器307可以包括以下详细描述的一个或多个组件,其被实现为可执行文件、库、静态库、函数或任何其他可执行组件。

[0063] 在一些示例性实施例中,桶确定器310可以被配置为基于所标记的数据集的标签来确定对于所标记的数据集的桶。桶可以是保持实例、对实例的引用等的数据结构。每个桶可以保持具有相同标签的实例或具有相似标签的实例。

[0064] 在一些示例性实施例中,聚类确定器320可以被配置为确定桶内的实例的聚类。在一些示例性实施例中,聚类确定器320可以使用任何聚类技术来确定聚类,例如但不限于k均值聚类、基于质心的聚类、基于分布的聚类、具有噪声的应用的基于密度的空间聚类(DBSCAN)、使用分层的平衡迭代减少和聚类(BIRCH)等。聚类确定器320可以利用任何聚类算法,例如但不限于K均值、分层聚类、仿射传播等。在一些示例性实施例中,聚类确定器320可以独立地确定由桶确定器310确定的每个桶中的聚类。

[0065] 在一些示例性实施例中,替选特征确定器330可以被配置为确定替选特征集合。替选特征集合可以是特征集合的泛化。替选特征集合可以包括具有与具有域的特征相对应的泛化域的泛化特征。由替选特征集合包括的每个泛化特征可以对应于聚类。

[0066] 在一些示例性实施例中,泛化实例确定器340可以被配置为确定具体实例的泛化实例。在一些示例性实施例中,泛化实例可以具有对替选特征集合的评估。泛化实例确定器340可以接收具有对特征的评估的实例。基于评估并且基于替选特征集合,泛化实例确定器340可以输出泛化实例。

[0067] 在一些示例性实施例中,具体实例确定器350可以被配置为基于泛化实例来确定具体实例。在一些示例性实施例中,泛化实例可以指示它代表作为聚类中的成员的实例。具体实例确定器350可以将聚类的另一成员确定为具体实例。在一些示例性实施例中,可以将具体实例确定为统计表示,例如由聚类包括的实例的模式、均值等,聚类的质心等。附加地或替选地,可以在由聚类包括的所观测的实例之间选择具体实例确定器350。附加地或替选地,在泛化实例指示多个聚类的情况下,可以选择最顶层的聚类并且可以基于最顶层的聚类生成具体实例。附加地或替选地,可以基于多个聚类来生成具体实例,诸如可以通过生成作为所有多个聚类的成员的实例来生成具体实例。

[0068] 在一些示例性实施例中,标签确定器360可以被配置为确定具体实例的标签。标签

确定器360可以例如通过应用决策模型来确定具体实例的标签。在一些示例性实施例中,确定标签可以在服务器上执行。计算机化的边缘设备可以被配置为获得实例并且确定该实例的泛化实例。计算机化的边缘设备可以将泛化实例、具体实例等传送到服务器,以便服务器基于泛化实例确定实例的标签。在这样的实施例中,最小化的数据被传输到服务器,从而减少了本地存储在边缘设备上的潜在私有信息的暴露。

[0069] 在一些示例性实施例中,服务器可以通过利用预测模型来确定标签。可以基于所标记的数据集来训练预测模型。附加地或替选地,服务器可以将预测模型传送到边缘设备。在一些示例性实施例中,边缘设备可以被配置为在本地基于第二实例确定实例的标签,而不传送任何数据。

[0070] 在一些示例性实施例中,将最终用户暴露给基于决策模型确定的映射函数、基于决策模型确定的计算范围、基于决策模型确定的聚类等可能会公开关于决策模型本身、关于其训练数据等的一些信息。在一些情况下,可能期望防止最终用户、客户端设备等获得对关于决策模型的这样的信息(这可能是商业秘密)的访问。

[0071] 在一些示例性实施例中,函数加密可以使得能够加密对某个函数的输入,并且通过加密数据来计算函数的结果,而执行计算的一方只能解密计算的结果。在一些示例性实施例中,用于给定函数f的函数加密方案包括以下四种算法:

[0072] • $(pk,msk)\leftarrow setup(1^{\lambda})$:创建公共密钥pk和主密钥msk。

[0073] • $sk\leftarrow Keygen(m_sk,f)$:使用主密钥为函数f生成新密钥sk。

[0074] • $c\leftarrow Enc(pk,x)$:使用公共密钥对消息x进行加密。

[0075] • $y\leftarrow Dec(sk,_c)$:使用密钥计算y=f(x),其中x是c加密的值。

[0076] 在一些示例性实施例中,可以使用函数加密方案来计算函数f,例如将原始数据点映射到对应的簇或范围的函数。在一些示例性实施例中,边缘设备(例如由用户使用的客户终端)可以从用户获得数据(d)。可以使用公共密钥(pk)对数据(d)进行加密。边缘设备可能会将加密的数据传输到服务器,从而阻止服务器访问数据本身。服务器可以使用密钥(sk)来解密将函数f应用于加密的数据(f(d))的结果,以接收与用户相关的关联的聚类或特征范围。在一些示例性实施例中,使用函数加密,客户端设备和用户可能无法访问使得能够进行逆向工程或理解决策模型(或用来训练决策模型的训练数据集)的信息,同时仍然保护其隐私并且确保实际数据点(d)没有被公开给服务器。

[0077] 本发明可以是系统、方法和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质,其上载有用于使处理器实现本发明的各个方面的计算机可读程序指令。

[0078] 计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是——但不限于——电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括:便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式压缩盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身,诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波(例如,通过光纤电缆的光脉冲)、或者通过电线传输

的电信号。

[0079] 此处所描述的计算机微观程序指令可以从计算机微观存储介质下载到各个计算/处理设备,或者通过网络,例如串行,重定向,广域网和/或无线网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,光纤传输,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替代程序指令,并复制该计算机精细程序指令,以供存储在各个计算/处理设备中的计算机细分存储介质中。

[0080] 用于执行本发明操作的计算机程序指令可以是汇编指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据或者以一种或多种编程语言的任意组合编写的源代码或目标代码,所述编程语言包括面向对象的编程语言—诸如Java、Smalltalk、C++等,以及传统过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上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。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。在一些实施例中,通过利用计算机可读程序指令的状态信息来个性化定制电子电路,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),该电子电路可以执行计算机可读程序指令,从而实现本发明的各个方面。

[0081] 这里参照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/或框图描述了本发明的各个方面。应当理解,流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合,都可以由计算机可读程序指令实现。

[0082] 这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器,从而生产出一种机器,使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时,产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中,这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作,从而,存储有指令的计算机可读介质则包括一个制造品,其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0083] 也可以把计算机精确的程序指令加载到计算机,其他数据处理装置,或其他设备上,用来在计算机,其他数据处理装置或其他设备上执行各种操作步骤,以产生计算机实现的过程,从而由此在计算机,其他异步数据处理装置,或其他设备上执行的指令实现流程图和/或提示中的一个或多个方框中规定的功能/动作。

[0084] 附图中的流程图和框图显示了根据本发明的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分,所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个连续的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的是,框图和/或

流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0085] 本文中使用的术语仅出于描述特定实施例的目的，而不是旨在限制本发明。如本文中所使用的，单数形式“一”、“一个”和“该”也旨在包括复数形式，除非上下文另外明确地指出。将进一步理解的是，术语“包括”和/或“包含”当在本说明书中使用时指定了所述特征、整体、步骤、操作、元素和/或组件的存在，但是并不排除一个或多个其他特征、整体、步骤、操作、元素、组件和/或其组的存在或添加。

[0086] 以下权利要求中的所有装置或步骤加上功能元件的对应结构、材料、动作和等同物旨在包括用于与具体要求保护的其他要求保护的元件组合地执行功能的任何结构、材料或动作。已经出于说明和描述的目的给出了本发明的描述，但是并不旨在是穷举的或将本发明限制为所公开的形式。在不偏离本发明的范围和精神的情况下，许多修改和变型对于本领域普通技术人员将是明显的。选择和描述实施例是为了最好地解释本发明的原理和实际应用，并且使得本领域的其他普通技术人员能够理解本发明的各种实施例，这些实施例具有适于所设想的特定用途的各种修改。

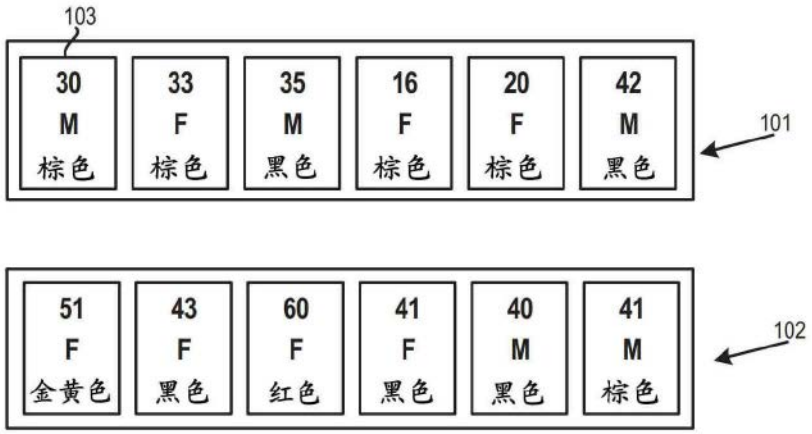

103
30 M 棕色
33 F 棕色
35 M 黑色
16 F 棕色
20 F 棕色
42 M 黑色
101
51 F 金黄色
43 F 黑色
60 F 红色
41 F 黑色
40 M 黑色
41 M 棕色
102

图1A

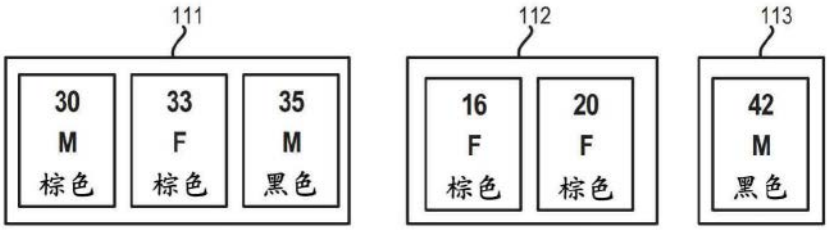

111
112
113
30 M 棕色
33 F 棕色
35 M 黑色
16 F 棕色
20 F 棕色
42 M 黑色

图1B

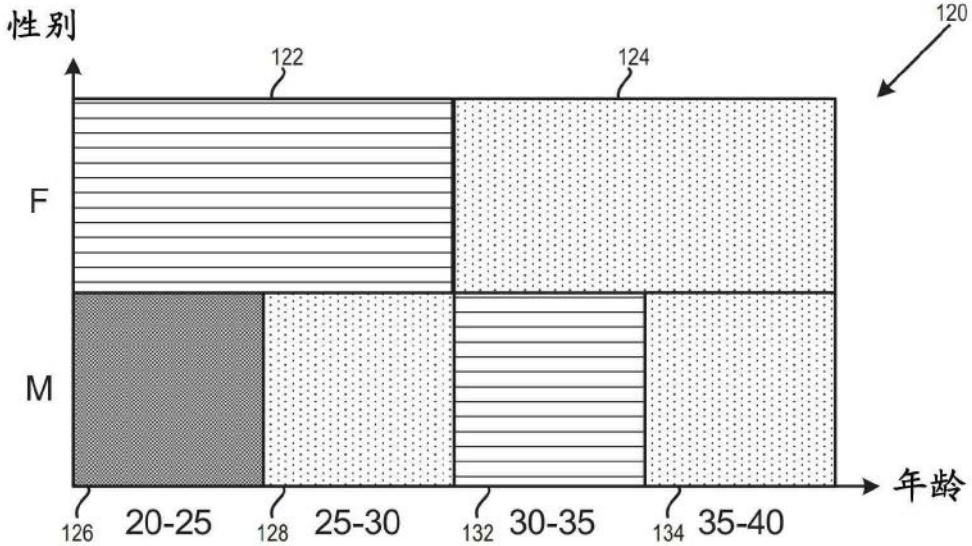

性别
120
122
124
F
M
年龄
126
20-25
128
25-30
132
30-35
134
35-40

图1C

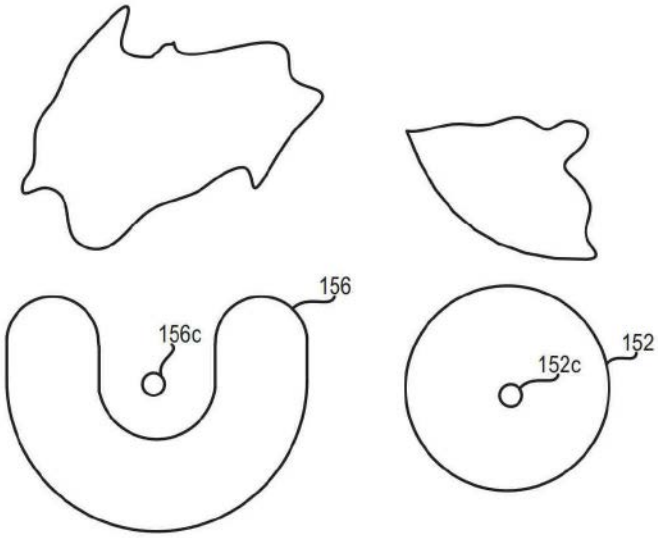


图1D

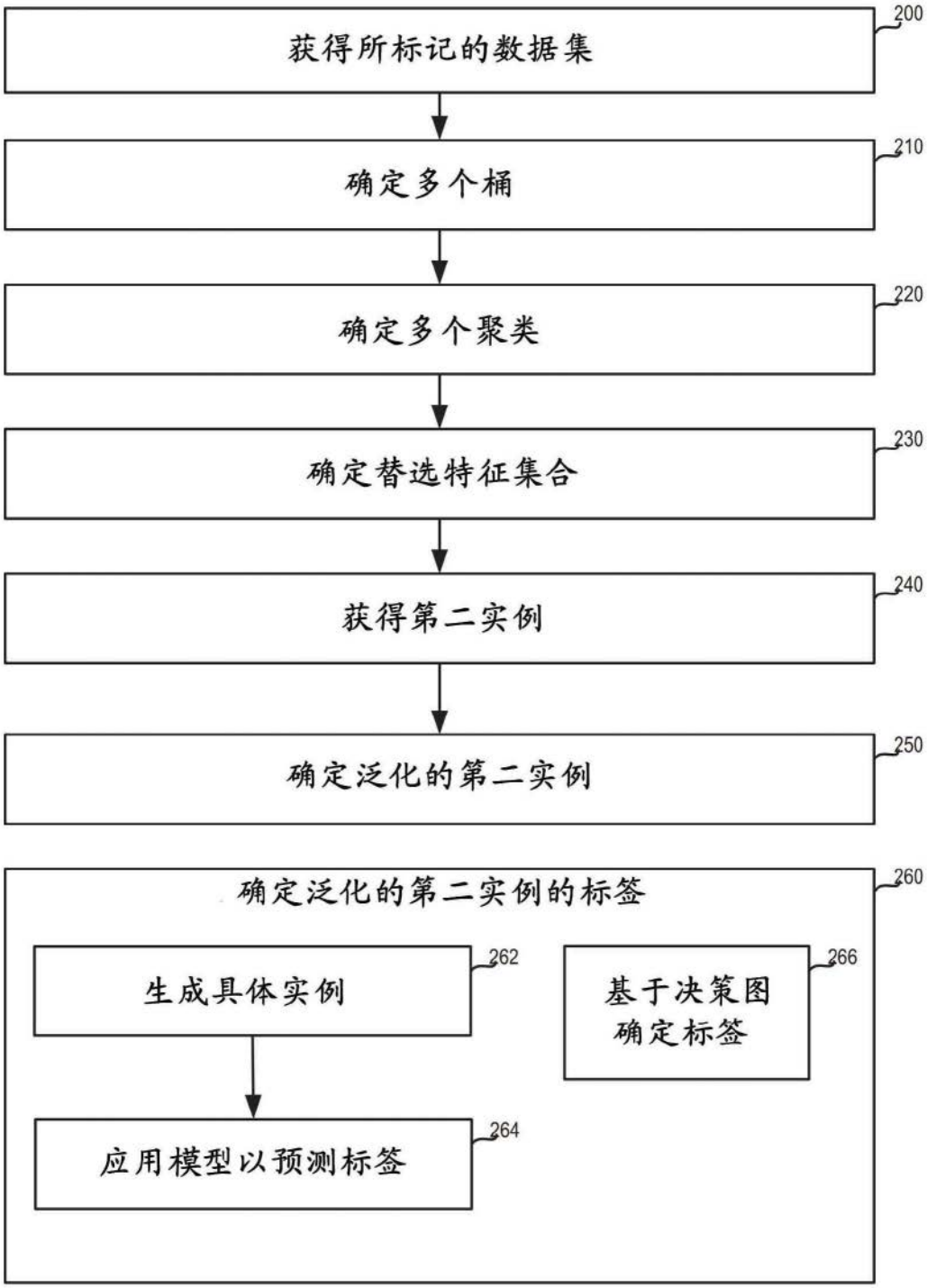

获得所标记的数据集
200
确定多个桶
210
确定多个聚类
220
确定替选特征集合
230
获得第二实例
240
确定泛化的第二实例
250
确定泛化的第二实例的标签
260
生成具体实例
262
应用模型以预测标签
264
基于决策图
确定标签
266

图2

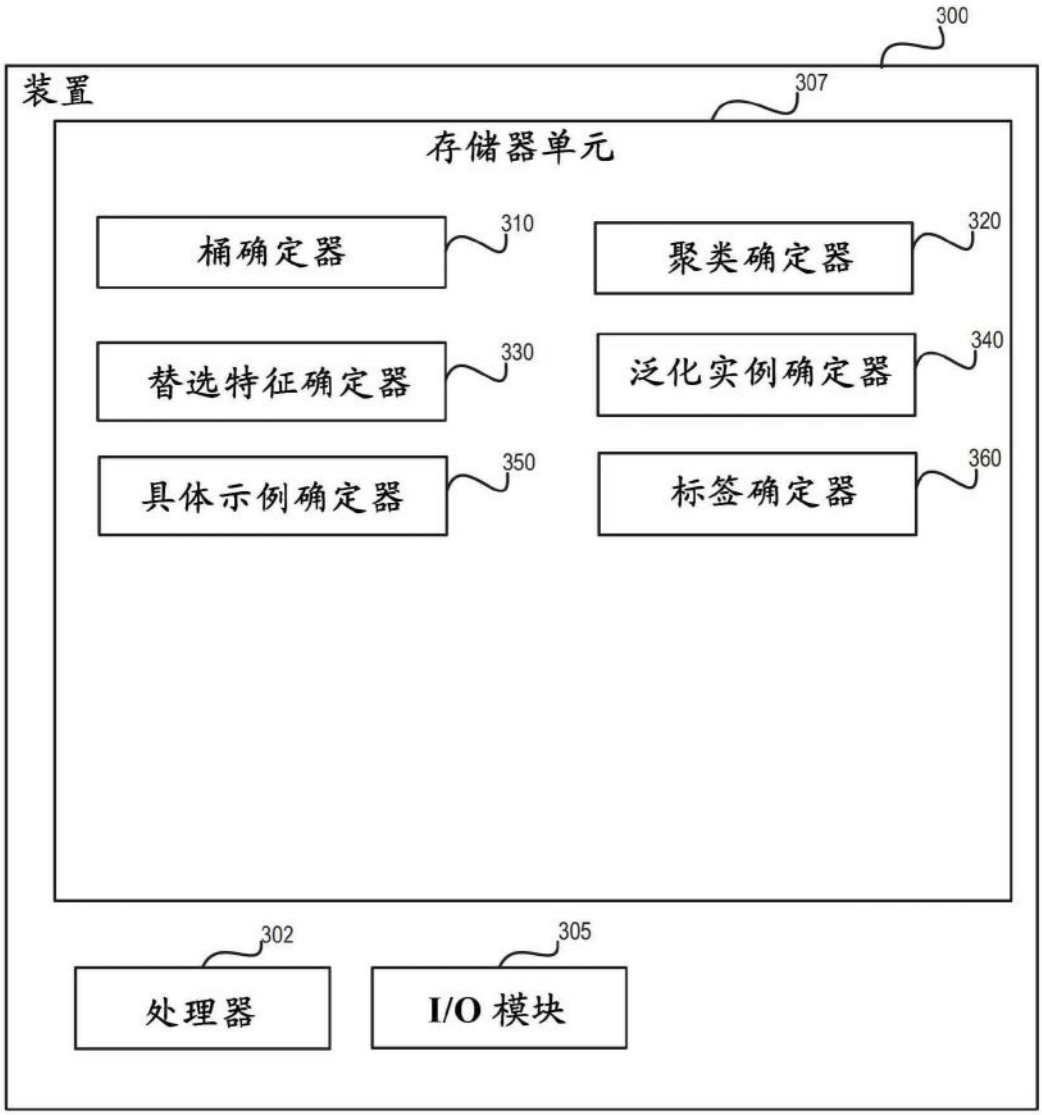

300
装置
307
存储器单元
桶确定器
310
聚类确定器
320
替选特征确定器
330
泛化实例确定器
340
具体示例确定器
350
标签确定器
360
302
处理器
305
I/O 模块

图3