



## (12)发明专利申请

(10)申请公布号 CN 109242106 A

(43)申请公布日 2019.01.18

(21)申请号 201811043185.9

(22)申请日 2018.09.07

(71)申请人 百度在线网络技术(北京)有限公司

地址 100080 北京市海淀区上地十街10号

(72)发明人 赵岷 程健一 秦华鹏

(74)专利代理机构 北京市金杜律师事务所

11256

代理人 李辉

(51)Int.Cl.

G06N 20/00(2019.01)

G06K 9/62(2006.01)

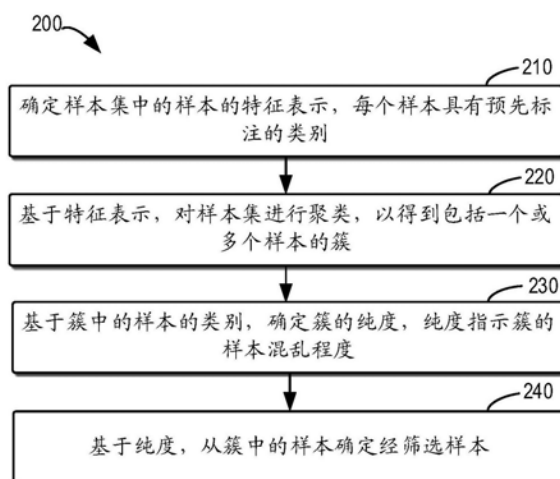
权利要求书3页 说明书10页 附图4页

### (54)发明名称

样本处理方法、装置、设备和存储介质

### (57)摘要

根据本公开的示例实施例,提供了样本处理方法、装置、设备和计算机可读存储介质。样本处理方法包括确定样本集中的样本的特征表示,每个样本具有预先标注的类别。该方法还包括基于特征表示,对样本集进行聚类,以得到包括一个或多个样本的簇。该方法进一步包括基于簇中的样本的类别,确定簇的纯度,纯度指示簇的样本混乱程度。该方法进一步包括基于纯度,从簇中的样本确定经筛选样本。以此方式,能够实现对整个样本集的筛选,从而获得高质量的标注样本。



1. 一种样本处理方法, 包括:  
确定样本集中的样本的特征表示, 每个样本具有预先标注的类别;  
基于所述特征表示, 对所述样本集进行聚类, 以得到包括一个或多个样本的簇;  
基于所述簇中的样本的类别, 确定所述簇的纯度, 所述纯度指示所述簇的样本混乱程度; 以及  
基于所述纯度, 从所述簇中的样本确定经筛选样本。
2. 根据权利要求1所述的方法, 其中从所述簇中的样本确定经筛选样本包括:  
响应于所述纯度高于阈值纯度, 基于所述簇中的样本的类别来确定所述经筛选样本。
3. 根据权利要求2所述的方法, 其中确定所述经筛选样本包括:  
响应于所述簇中的样本的类别相同, 将所述簇中的样本确定为所述经筛选样本。
4. 根据权利要求2所述的方法, 其中确定所述经筛选样本包括:  
响应于所述簇中的样本具有多种类别, 确定每种类别的样本数目;  
基于每种类别的样本数目, 确定所述簇中具有最大样本数目的类别; 以及  
将具有确定的所述类别的样本确定为所述经筛选样本。
5. 根据权利要求1所述的方法, 其中从所述簇中的样本确定经筛选样本包括:  
响应于所述纯度低于阈值纯度, 确定所述簇中的样本数目占所述样本集中的样本总数目的比例;  
响应于所述比例超过阈值比例, 对所述簇中的样本进行聚类, 以得到聚类结果; 以及  
至少基于所述聚类结果, 将所述簇中的样本中的至少一部分确定为经筛选样本。
6. 根据权利要求1所述的方法, 其中确定所述特征表示包括:  
将所述样本集应用于特征提取模型, 以得到与所述样本集相关联的隐层神经元; 以及  
基于所述隐层神经元, 确定所述样本集中的样本的特征表示。
7. 根据权利要求6所述的方法, 还包括:  
至少基于所述经筛选样本, 确定所述样本集的一个子集, 所述子集包括从与所述样本集相关联的至少一个簇得到的经筛选样本;  
将所述子集应用于所述特征提取模型, 以得到所述子集中的样本的更新特征表示; 以及  
基于所述更新特征表示, 对所述子集进行聚类, 以基于聚类结果来更新所述经筛选样本。
8. 根据权利要求1所述的方法, 其中确定所述特征表示包括:  
确定所述样本集中的样本在预定义的特征空间中的特征值, 作为所述特征表示。
9. 根据权利要求8所述的方法, 还包括:  
至少基于所述经筛选样本, 确定所述样本集的一个子集, 所述子集包括从与所述样本集相关联的至少一个簇得到的经筛选样本; 以及  
基于所述特征表示, 对所述子集进行聚类, 以基于聚类结果来更新所述经筛选样本。
10. 根据权利要求1所述的方法, 其中确定所述簇的所述纯度包括:  
确定所述簇中每种类别的样本数目;  
基于每种类别的样本数目, 确定最大样本数目; 以及  
基于所述最大样本数目和所述簇中的样本总数目确定所述纯度。

11. 一种样本处理装置, 包括:

第一表示确定模块, 被配置为确定样本集中的样本的特征表示, 每个样本具有预先标注的类别;

第一聚类模块, 被配置为基于所述特征表示, 对所述样本集进行聚类, 以得到包括一个或多个样本的簇;

第一纯度确定模块, 被配置为基于所述簇中的样本的类别, 确定所述簇的纯度, 所述纯度指示所述簇的样本混乱程度; 以及

样本筛选模块, 被配置为基于所述纯度, 从所述簇中的样本确定经筛选样本。

12. 根据权利要求11所述的装置, 其中所述样本筛选模块包括:

第一样本确定模块, 被配置为响应于所述纯度高于阈值纯度, 基于所述簇中的样本的类别来确定所述经筛选样本。

13. 根据权利要求12所述的装置, 其中所述第一样本确定模块包括:

第二样本确定模块, 被配置为响应于所述簇中的样本的类别相同, 将所述簇中的样本确定为所述经筛选样本。

14. 根据权利要求12所述的装置, 其中所述第一样本确定模块包括:

第一数目确定模块, 被配置为响应于所述簇中的样本具有多种类别, 确定每种类别的样本数目;

最大类别确定模块, 被配置为基于每种类别的样本数目, 确定所述簇中具有最大样本数目的类别; 以及

第三样本确定模块, 被配置为将具有确定的所述类别的样本确定为所述经筛选样本。

15. 根据权利要求11所述的装置, 其中所述样本筛选模块包括:

样本比例确定模块, 被配置为响应于所述纯度低于阈值纯度, 确定所述簇中的样本数目占所述样本集中的样本总数目的比例;

第二聚类模块, 被配置为响应于所述比例超过阈值比例, 对所述簇中的样本进行聚类, 以得到聚类结果; 以及

第四样本确定模块, 被配置为至少基于所述聚类结果, 将所述簇中的样本中的至少一部分确定为经筛选样本。

16. 根据权利要求11所述的装置, 其中所述第一表示确定模块包括:

样本应用模块, 被配置为将所述样本集应用于特征提取模型, 以得到与所述样本集相关联的隐层神经元; 以及

第二表示确定模块, 被配置为基于所述隐层神经元, 确定所述样本集中的样本的特征表示。

17. 根据权利要求16中所述的装置, 还包括:

第一子集确定模块, 被配置为至少基于所述经筛选样本, 确定所述样本集的一个子集, 所述子集包括从与所述样本集相关联的至少一个簇得到的经筛选样本;

第一子集应用模块, 被配置为将所述子集应用于所述特征提取模型, 以得到所述子集中的样本的更新特征表示;

第一样本更新模块, 被配置为基于所述更新特征表示, 对所述子集进行聚类, 以基于聚类结果来更新所述经筛选样本。

18. 根据权利要求11所述的装置,其中所述第一表示确定模块包括:

第三表示确定模块,被配置为确定所述样本集中的样本在预定义的特征空间中的特征值,作为所述特征表示。

19. 根据权利要求18中所述的装置,还包括:

第二子集确定模块,被配置为至少基于所述经筛选样本,确定所述样本集的一个子集,所述子集包括从与所述样本集相关联的至少一个簇得到的经筛选样本;以及

第二样本更新模块,被配置为基于所述特征表示,对所述子集进行聚类,以基于聚类结果来更新所述经筛选样本。

20. 根据权利要求11所述的装置,其中所述第一纯度确定模块包括:

第二数目确定模块,被配置为确定所述簇中每种类别的样本数目;

最大数目确定模块,被配置为基于每种类别的样本数目,确定最大样本数目;以及

第二纯度确定模块,被配置为基于所述最大样本数目和所述簇中的样本总数目确定所述纯度。

21. 一种设备,所述设备包括:

一个或多个处理器;以及

存储装置,用于存储一个或多个程序,当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现如权利要求1-10中任一项所述的方法。

22. 一种计算机可读存储介质,其上存储有计算机程序,所述程序被处理器执行时实现如权利要求1-10中任一项所述的方法。

## 样本处理方法、装置、设备和存储介质

### 技术领域

[0001] 本公开的实施例主要涉及计算机领域,并且更具体地,涉及样本处理方法、装置、设备和计算机可读存储介质。

### 背景技术

[0002] 随着人工智能的普及,机器学习技术得到越来越广泛的应用。机器学习使用统计技术来赋予计算机系统利用数据“学习”(例如,逐步提高特定任务的表现)的能力。监督学习是一种类型的机器学习任务,其基于示例输入-输出对学习将输入映射到输出的函数。在监督学习中,从由一组训练示例组成的经标注的训练数据(即,标注样本)中推断出这样的函数。

[0003] 监督学习任务通常需要大规模高精度的标注样本,并且标注样本的质量影响着监督学习的学习效果。由于人工标注样本的代价大且效率低,因此很多应用采取机器自动标注样本的方式。机器自动标注的样本虽然可以达到很大规模,但往往精度远低于人工标注,这制约了机器学习模型的训练效果。因此,需要高效的样本清洗和筛选方法,以提高用于特别是监督学习的机器学习的标注样本的质量。

### 发明内容

[0004] 根据本公开的示例实施例,提供了一种样本处理方案。

[0005] 在本公开的第一方面中,提供了一种样本处理方法。该方法包括确定样本集中的样本的特征表示,每个样本具有预先标注的类别。该方法还包括基于特征表示,对样本集进行聚类,以得到包括一个或多个样本的簇。该方法进一步包括基于簇中的样本的类别,确定簇的纯度,纯度指示簇的样本混乱程度。该方法进一步包括基于纯度,从簇中的样本确定经筛选样本。

[0006] 在本公开的第二方面中,提供了一种样本处理装置。该装置包括第一表示确定模块,被配置为确定样本集中的样本的特征表示,每个样本具有预先标注的类别。该装置还包括第一聚类模块,被配置为基于特征表示,对样本集进行聚类,以得到包括一个或多个样本的簇。该装置进一步包括第一纯度确定模块,被配置为基于簇中的样本的类别,确定簇的纯度,纯度指示簇的样本混乱程度。该装置进一步包括样本筛选模块,被配置为基于纯度,从簇中的样本确定经筛选样本。

[0007] 在本公开的第三方面中,提供了一种设备,包括一个或多个处理器;以及存储装置,用于存储一个或多个程序,当一个或多个程序被一个或多个处理器执行,使得一个或多个处理器实现根据本公开的第一方面的方法。

[0008] 在本公开的第四方面中,提供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根据本公开的第一方面的方法。

[0009] 应当理解,发明内容部分中所描述的内容并非旨在限定本公开的实施例的关键或重要特征,亦非用于限制本公开的范围。本公开的其它特征将通过以下的描述变得容易理

解。

## 附图说明

[0010] 结合附图并参考以下详细说明,本公开各实施例的上述和其他特征、优点及方面将变得更加明显。在附图中,相同或相似的附图标注表示相同或相似的元素,其中:

[0011] 图1示出了本公开的多个实施例能够在其中实现的示例环境的示意图;

[0012] 图2示出了根据本公开的实施例的样本处理过程的流程图;

[0013] 图3示出了根据本公开的一些实施例的用于确定特征表示的神经网络的示意图;

[0014] 图4示出了根据本公开的一些实施例的聚类结果的示意图;

[0015] 图5示出了根据本公开的一些实施例的确定经筛选样本的过程的流程图;

[0016] 图6示出了根据本公开的多个实施例的样本处理装置的示意性框图;以及

[0017] 图7示出了能够实施本公开的多个实施例的计算设备的框图。

## 具体实施方式

[0018] 下面将参照附图更详细地描述本公开的实施例。虽然附图中显示了本公开的某些实施例,然而应当理解的是,本公开可以通过各种形式来实现,而且不应该被解释为限于这里阐述的实施例,相反提供这些实施例是为了更加透彻和完整地理解本公开。应当理解的是,本公开的附图及实施例仅用于示例性作用,并非用于限制本公开的保护范围。

[0019] 在本公开的实施例的描述中,术语“包括”及其类似用语应当理解为开放性包含,即“包括但不限于”。术语“基于”应当理解为“至少部分地基于”。术语“一个实施例”或“该实施例”应当理解为“至少一个实施例”。术语“第一”、“第二”等等可以指代不同的或相同的对象。下文还可能包括其他明确的和隐含的定义。

[0020] 在本公开的实施例的描述中,如本领域技术人员所理解的,术语“聚类”是指试图将样本集中的样本划分为通常是不相交的子集的过程,其中所得到的每个子集被称为一个“簇”。每个簇可能对应于一些潜在的概念。需要说明的是,尽管在本文中样本集中的样本具有预先标注的类别,但是聚类过程并不使用这种类别信息。另外,在一些情况下还可以对得到的簇(可以被视为另一样本集)再进行聚类,以得到包括该簇中的样本中的一个或多个样本的子集(也可以被称为子簇)。

[0021] 在本公开的实施例的描述中,如本领域技术人员所理解的,术语“神经网络”是指由具有适应性的简单单元组成的广泛并行互连的网络,它的组织能够模拟生物神经系统对真实世界物体所做出的交互反应。在神经网络中,最基本的成分是“神经元”模型,即上述定义中的“简单单元”。

[0022] 传统上,如前文所提及的,主要存在两种对样本数据进行“清洗”或筛选的方案,即基于人工规则的样本筛选和基于半监督学习的样本筛选。在基于人工规则的方案中,需要人工发现错误样本的规律,并构造错误样本的模板,通过模式匹配的方式去除错误或噪音或样本,从而得到经筛选的样本。这种方案规则构造成本高且适用范围有限,仅适用于有显式规律并能用模板表示的样本。

[0023] 在基于半监督学习的方案中,通过预先选择的少量高质量样本来训练机器学习模型,再使用训练得到的模型预测全量样本,选择预测置信度高的样本增加到高质样本集合

中。这种方案一方面依赖于初始训练样本的质量,另一方面选择的样本偏向拟合初始训练样本,从而导致难以覆盖整个样本空间。

[0024] 根据本公开的实施例,提出了一种样本处理方案,用于从具有预先标注的类别的全量样本中筛选优质样本。在该方案中,基于样本的特征表示对样本集中的样本进行聚类,并且基于样本的类别,确定聚类得到的每个簇的纯度。针对每个簇,基于纯度采取不同的后处理策略来确定经筛选样本。以此方式,能够从有噪音的全量样本中筛选出优质样本,以供后续监督学习任务使用。因此,本公开的方案能够有利地实现自动、有效且低成本的样本筛选。

[0025] 以下将参照附图来具体描述本公开的实施例。

[0026] 图1示出了本公开的多个实施例能够在其中实现的示例环境100的示意图。在该示例环境100中,由计算设备102对样本集101进行处理,以从中确定经筛选样本,例如以从中筛选优质样本。应当理解,图1中示出的环境100仅是示例性的,还可以使用多个计算设备来确定处理样本集101。还应当理解,计算设备102可以是固定式计算设备,也可以是便携式计算设备,诸如移动电话、平板计算机等。

[0027] 样本集101(在本文中有时也称为初始样本集101)包括多个样本。如图1所示,样本集101可以包括样本110-1至110-9(为便于描述,统称为样本110)。应当理解,尽管图1示出了样本集101中的九个样本110,但是样本集101可以包括更多或更少的样本110。在一些实施例中,样本集101可以包括大规律的样本110,例如,几千至几万个样本。

[0028] 样本110可以是各种类型的样本,包括但不限于文本、图像、视频、音频等。例如,样本110-1至110-9可以分别是一篇或多篇文章、一段或多段文字、一个或多个语句。样本110-1至110-9可以分别是一个或多个图像、一段或多段视频。本公开的实施例在此方面不受限制。

[0029] 样本110具有预先标注(或标记)的类别,例如,如图1所示的类别A、B和C。在图1的示例中,样本110-1至110-4具有类别A,样本110-5至110-7具有类别B,并且样本110-8和110-9具有类别C。应当理解,图1中示出样本110具有三种类别A、B和C仅仅是示意性的,而无意限制本公开的范围。本公开的实施例可以应用于各种分类情形,诸如二分类、多分类情形。

[0030] 在本文中,类别可以用于指示样本在某一方面具有相同或相似的属性。仅举例而言,在样本110是文章的情况下,类别可以表示文章的类型。例如,类别A、B和C指示对应的样本分别被标注为新闻类文章、评论类文章和科普类文章。在样本110是图像的情况下,类别可以表示图像中包含的对象类型。例如,类别A、B和C指示对应的样本分别被标注为包含人、动物和植物。类别可以根据需要指示样本的各种相同或相似属性,本公开的范围在此方面不受限制。

[0031] 样本110的类别A、B和C可以通过各种方法标注的。例如,可以由人工标注。样本110也可以是通过数据挖掘获得的、具有预定类别的样本。样本110的类别还可以是由其他模型或系统生成的。本公开的范围在此方面不受限制。

[0032] 一般而言,这些预先标注的类别可能不是完全准确的,这导致了样本集101的噪音,即存在噪音样本。举例而言,图1中被预先标注为类别B的样本110-7实际上应当为类别C,那么这样的样本110-7即为噪音样本或错误样本,需要被去除或被纠正。

[0033] 计算设备102可以利用本文中公开的方法对样本集101进行处理,以从样本110中确定至少部分样本作为经筛选样本。例如,如图1所示,计算设备102可以从样本集101中确定包括样本110-1、110-2、110-5、110-6和110-9在内的样本作为经筛选样本。这些经筛选样本110-1、110-2、110-5、110-6和110-9可以构成例如经筛选样本集103,其是样本集101的子集。经筛选样本在本文中用于指代从样本集中选择的或以其他方式确定的样本。

[0034] 应当理解,尽管图1示出样本110-1、110-2、110-5、110-6和110-9为经筛选样本,但是经筛选样本可以包括更多或更少的样本110。在一些实施例中,计算设备102还可以利用本文中公开的方法对经筛选样本集103进行处理,以进一步筛选样本。

[0035] 为了更清楚地理解本公开的实施例所提供的样本处理方案,将参照图2来进一步描述本公开的实施例。图2示出了根据本公开的实施例的样本处理过程200的流程图。过程200可以由图1的计算设备102来实现。为便于讨论,将结合图1来描述过程200。

[0036] 在210,计算设备102确定样本集101中的样本110的特征表示。每个样本110均具有预先标注的类别,例如,样本110-1至110-4具有类别A,样本110-5至110-7具有类别B,并且样本110-8和110-9具有类别C。特征表示在本文中用于指示与样本110的一个或多个属性相关联的特征的子集,其可以抽象地或以数学方式来描述样本110。特征表示通常为多维向量或矩阵。计算设备102可以使用任何能够将初始的样本110映射为特征空间中的特征向量的方法来确定特征表示。

[0037] 在一些实施例中,可以使用预定义的特征空间。计算设备102可以确定样本集101中的样本110在预定义的特征空间中的特征值,并且将这些特征值构成的特征向量作为特征表示。例如,在样本110是文本的情况下,如果特征空间是字典中的每个字,则样本110的特征表示可以为字向量;如果特征空间是词典中的每个词,则样本110的特征表示可以为词向量。

[0038] 在一些实施例中,可以使用机器学习方法学习样本的特征表示。计算设备102可以使用特征提取模型来确定特征表示,该特征提取模型可以基于任何机器学习方法。例如,特征提取模型可以包括神经网络,诸如卷积神经网络CNN、递归神经网络等。

[0039] 计算设备102可以将样本集101应用于特征提取模型,即应用于神经网络,以得到与样本集101中的样本110相关联的隐层(hidden layer)神经元,然后计算设备102可以基于所得到的隐层神经元来确定样本集101中的样本110的特征表示。即,计算设备102可以将生成的隐层向量作为样本110的特征表示。例如,在样本110是文本数据的情况下,可以使用卷积神经网络CNN分类器进行训练,将模型训练过程中生成的隐层向量输出作为样本的特征向量。

[0040] 下面结合图3来描述如何通过神经网络确定样本110的特征表示。图3示意性地图了根据本公开的一些实施例的用于确定特征表示的神经网络300。神经网络300可以包括输入层310,其包括输入层神经元311、312和313。输入层310的各个神经元311、312和313接收样本110中的每个样本的相应特征值作为输入。包括隐层神经元321、322、323和324的隐层320可以接收来自输入层310的输出,并经过一定的运算向下一层330输出。包括神经元331、332和333的层330可以是输出层,也可以是下一隐层。

[0041] 计算设备102可以例如基于隐层神经元321、322、323和324来确定样本110的特征表示。计算设备102可以将隐层神经元321、322、323和324的输出值确定为特征向量在相应

维度上的取值,并且将该特征向量作为特征表示。在图3的示例中,用作特征表示的特征向量具有4个维度。在层330是隐层的情况下,也可以基于神经元331、332和333来确定特征表示。

[0042] 应当理解,图3中所示的神经网络300仅仅是示意性的,而无意限制本公开的范围。计算设备102可以使用具有任何数目的隐层和任何数目的隐层神经元的神经网络。另外,计算设备102可以根据实际需要(例如,样本类型、任务需求)来确定使用哪一隐层的神经元。

[0043] 在这样的实施例中,特征表示是由可训练的神经网络生成的隐层数据确定的,而不是对样本特征的直接表示。这样的特征表示可以呈现与目标联系更为密切的表示,有利于后续的聚类。此外,应当注意,在这样的实施例中,由于使用神经网络的目标是获取样本110的隐层向量表示,因此对神经网络模型的分类精度没有严格要求,可以直接使用全量的噪音样本进行训练。

[0044] 继续参考图2,在220,计算设备102基于在210确定的特征表示,对样本集101进行聚类,从而得到包括样本110中的一个或多个样本的簇。聚类可以使用任何合适的聚类算法,例如,k均值(K-Means)聚类算法、诸如DBSCAN的密度聚类算法等。

[0045] 聚类结果通常为n个簇,每个簇通常包括不同数目的样本110。图4示出了根据本公开的一些实施例的聚类结果400的示意图。图4示意性地示出了对样本集101进行聚类而得到的簇401、402和403。簇401中至少包含样本110-1和110-2;簇402中至少包含样本110-3、110-5、110-6和110-8;簇403中至少包含样本110-4、110-7和110-9。应当理解,图4中所示的聚类结果400仅仅是示意性的,而无意限制本公开的范围。聚类结果400还可以包括更多或更少的簇,并且簇401、402和403还可以包含更多或更少的样本110。

[0046] 继续参考图2,在230,计算设备102基于簇中的样本的类别确定簇的纯度,纯度用于指示簇的样本混乱程度。簇中样本的类别越多且各个类别的样本数目越接近,簇的纯度越低;反之,簇中某一类别的样本占比越高,簇的纯度越高。纯度可以使用在0至1之间的数字、或者百分比来表示,也可以使用任何合适的数字来表示。计算设备102可以针对簇401、402和403分别确定纯度。可以根据不同应用需求来设计不同的纯度计算方法。

[0047] 在一些实施例中,可以使用簇中最大数目类别的样本占整个簇的样本比例作为簇的纯度。以图4中所示的簇402为例,计算设备102可以确定簇402中每种类别(即,类别A、B和C)的样本数目。例如,计算设备102可以确定:在簇402中,类别A的样本数目为30,类别B的样本数目为60,并且类别C的样本数目为10。

[0048] 然后,计算设备102可以基于类别A、B和C的样本数目确定最大样本数目。例如,计算设备102可以确定在簇402中类别B的样本最多,并且最大样本数目为60。计算设备102进而可以基于所确定的最大样本数目和簇402中的样本的总数目确定簇402的纯度。例如,簇402的样本总数目为100,则可以确定簇402的纯度为 $60/100=0.6$ 。对于聚类结果400中的其他簇(例如,簇401和403)可以使用相同的方式确定纯度。

[0049] 在一些实施例中,可以使用信息熵来确定簇的纯度。可以使用式(1)来计算每个簇401、402和403的信息熵H。

$$[0050] \quad H = -\sum_{i=1}^k p_i \log p_i \quad (1)$$

[0051] 其中 $p_i$ 表示第i类别的样本数目占簇的样本总数目的比例,k表示簇中的样本的总类别数。例如,对于图4中所示的簇402,k为3, $p_1$ 表示类别A的样本的比例0.6, $p_2$ 表示类别B的

样本的比例0.3,  $p_3$ 表示类别C的样本的比例0.1。

[0052] 应当注意,簇中的样本越混乱,相应的信息熵 $H$ 越大,因此可以使用信息熵的倒数 $1/H$ 作为簇402的纯度。对于聚类结果400中的其他簇(例如,簇401和403)可以使用相同的方式确定纯度。

[0053] 在240,计算设备102基于在230确定的纯度来从簇中的样本确定经筛选样本。计算设备102可以针对不同的纯度采用不同的后处理策略来从各个簇401、402和403中筛选优质样本。例如,可以基于纯度分为三种情况来采取不同的后处理策略。第一种情况是簇中仅包含一种类别(例如,类别A)的样本;第二种情况是簇中包含多种类别的样本,但是该簇的纯度较高(例如,高于阈值纯度);第三中情况是簇中包含多种类别的样本,但是该簇的纯度较低(例如,低于阈值纯度)。下文将参考图5来详细描述基于纯度来确定经筛选样本的过程。

[0054] 以上描述了根据本公开的实施例的样本处理过程200。以此方式,可以实现对整个样本集的过滤清洗,从而获得高质量的标注样本。因此,本公开的方案能够对样本进行清洗过滤,特别是能够对有噪音的大规模样本进行全量的清洗过滤,有效去除噪音样本,提升样本质量,从而有助于提高大规模有监督机器学习任务的效果。

[0055] 在一些实施例中,计算设备102还可以对所获得的经筛选样本重复图2中的过程200或过程200的一部分,以进一步筛选样本,从而获得更加优质的样本。具体而言,计算设备102可以基于从簇401、402、403中得到的经筛选样本,确定样本集102的一个子集。如前文关于图1所提及的,在本文中可以将这样的子集称为经筛选样本集,例如图1中所示的经筛选样本集103。经筛选样本集103可以包括从与样本集101相关的簇401、402、403得到的经筛选样本。

[0056] 在上文所描述的基于神经网络确定特征表示的实施例中,可以对经筛选样本集103重复图2中的框210-240,以更新经筛选样本。计算设备102可以将经筛选样本集103应用于框210中使用的特征提取模型(即,神经网络),以得到经筛选样本集103中的样本的更新特征表示。由于框210中使用的神经网络是可学习的,因此在应用与初始样本集101不同的经筛选样本集103时,对于同一样本(例如,样本110-1)可以得到不同于先前的特征表示,即得到更新特征表示。

[0057] 计算设备102然后可以基于更新特征表示,对经筛选样本集103进行聚类,以基于新的聚类结果(即,生成的簇及簇的纯度)来更新经筛选样本。即,计算设备102可以重复图2中的框230和240。

[0058] 在上文所描述的基于预定义的特征空间确定特征表示的实施例中,可以对经筛选样本集103重复图2中的框220-240,以更新经筛选样本。计算设备102可以基于先前在210确定的特征表示来对经筛选样本集103进行聚类。在这样的实施例中,由于特征表示是在预定义的特征空间中确定的,对于同一样本(例如,样本110-1)特征表示不变。

[0059] 与对初始样本集101的聚类相比,对经筛选样本集103进行聚类可以采用不同的聚类算法,或者采用不同的聚类参数(例如,聚类距离),或者两者的结合。计算设备102可以基于通过这样的方法获得的簇来进一步筛选样本,即更新经筛选样本。

[0060] 在这样情况下,通过筛选出的优质样本可以获得更好的聚类结果,通过更好的聚类结果来进一步筛选更优质的样本,直至满足终止条件为止。可以根据具体应用场景来设计终止条件。例如,终止条件可以是:在聚类结果中,低纯度簇的样本数目占样本集101的样

本总数目的比例小于预定阈值。终止条件还可以是经筛选的样本的数目小于预定数目。

[0061] 在这样的实施例中,通过迭代可以进一步筛选更优质的样本,从而提高最终获得的样本质量。以此方式,样本质量得以进一步提高,从而有利于提升后续监督学习任务的效果。

[0062] 如上文参考图2所提及的,计算设备102可以基于纯度来确定经筛选样本。下面将参考图4和图5来描述这一过程。图5示出了示出了根据本公开的一些实施例的确定经筛选样本的过程500的流程图。过程500可以视为图2中框240的实现。

[0063] 参考图5,在510,计算设备102可以确定簇的纯度是否高于阈值纯度。阈值纯度可以取决于具体应用场景,例如,样本110的类型、后续有监督学习任务对样本精度的要求等。如果计算设备102在510确定簇的纯度高于阈值纯度,则可以基于簇中的样本的类别来确定经筛选样本,即,过程500可以进行到框520。例如,对于簇401或簇402,计算设备102可以确定其纯度高于阈值纯度。

[0064] 在520,计算设备102可以确定簇401或簇402中的样本的类别是否相同。如果计算设备102在520确定簇中的样本的类别相同,则过程500可以进行到框530。例如,对于簇401,计算设备102可以在520确定簇401中的样本均为类别A。那么在530,计算设备102可以将簇401中包含的所有样本确定为经筛选样本,例如可以将簇401中的样本110-1、110-2等确定为经筛选样本。

[0065] 如果计算设备102在520确定簇中的样本的类别不同,即,簇中的样本具有多种类别,则过程500可以进行到框540。例如,对于簇402,计算设备102可以确定簇402中的样本具有多种类别A、B和C。在这种情况下,簇402具有多种类别的样本但纯度高于阈值纯度,这意味着在簇402中某种类别的样本占主导。

[0066] 在540,计算设备102可以确定簇402中每种类别的样本的数目。例如,如上文所描述的,计算设备102可以确定在簇402中类别A的样本数目为30、类别B的样本数目为60并且类别C的样本数目为10。在550,计算设备102可以基于每种类别的样本数目,确定簇402中具有最大样本数目的类别。例如,计算设备102可以确定在簇402中类别B的样本具有最大样本数目。

[0067] 在560,计算设备102可以将具有所确定的类别的样本确定为经筛选样本。例如,计算设备102可以将簇402中类别B的样本(诸如,样本110-5、110-6)确定为经筛选样本。

[0068] 对于除了类别B之外的其他类别的样本(诸如,样本110-3、110-8等),可以取决于不同的应用场景、任务需求等采取不同的处理。例如,如果其他类别的样本数目占样本集101的样本总数目的比例较小,则可以直接丢弃这些其他类别的样本。如果后续有监督学习任务对样本质量的要求较高,则可以将这些其他类别的样本输出,以进行人工标注。

[0069] 在一些情况下,还可以对簇402中的其他类别的样本进行自动纠错。例如,如果簇402中超过一定比例(例如,95%)的样本110均为类别B,那么可以将簇402中具有类别A和C的样本纠正为类别B,并且还可以将这样的经纠正样本也作为经筛选样本。

[0070] 继续参考框510,如果计算设备102在框510确定簇的纯度不高于阈值纯度,则可以基于簇中样本的数目来采取不同的处理,即,过程500可以进行到框570。例如,对于簇403,计算设备102可以在510确定簇403的纯度低于阈值纯度,这意味着簇403中各种类别的样本分布较为均匀。

[0071] 在570,计算设备102可以确定簇403中的样本数目占样本集101中的样本总数目的比例。在580,计算设备102可以确定该比例是否超过阈值比例,该阈值比例在本文也被称为上限阈值比例。如果确定该比例超过上限阈值比例,意味着簇403中的样本数目较大,则过程500可以进行到框590。

[0072] 在590,计算设备102可以对簇403中的样本(诸如,样本110-4、110-7、110-9等)进行再次聚类,以得到聚类结果。该聚类结果包括簇403中的样本110-4、110-7、110-9的一个或多个子集(也可以被称为一个或多个子簇)。在595,计算设备102可以基于该聚类结果,将簇403中的样本中的至少一部分确定为经筛选样本。例如,计算设备102可以对所得到的子簇重复图2中的框230和240来从中确定经筛选样本。

[0073] 如果计算设备102在580确定在570得到的比例不超过上限阈值比例,则计算设备102可以进一步取决于簇403中的样本数目来采取不同的处理。如果簇403中的样本数目较少,则可以丢弃簇403中的样本。例如,如果在570确定的簇403的样本比例小于另一阈值比例(为便于讨论,也称为下限阈值比例),则可以丢弃簇403中包含的全部样本。

[0074] 如果簇403中的样本数目适中,例如在570确定的簇403的样本比例大于下限阈值比例,则可以取决于具体应用场景而采取不同的处理。例如,对于样本精度要求高且样本总数目不足够大的情况,可以将簇403中的样本输出以进行人工标注。也可以对簇403中的样本进行小规模抽样,由人工判定后续处理。还可以将簇403中的样本全部丢弃,或者全部保留以留待下一轮迭代时优化。

[0075] 图6示出了根据本公开的实施例的用于对象确定的装置600的示意性框图。装置600可以被包括在图1的计算设备102中或者被实现为计算设备102。如图6所示,装置600包括第一表示确定模块610,被配置为确定样本集中的样本的特征表示,每个样本具有预先标注的类别。装置600还包括第一聚类模块620,被配置为基于特征表示,对样本集进行聚类,以得到包括一个或多个样本的簇。装置600进一步包括第一纯度确定模块630,被配置为基于簇中的样本的类别,确定簇的纯度,纯度指示簇的样本混乱程度。装置600进一步包括样本筛选模块640,被配置为基于纯度,从簇中的样本确定经筛选样本。

[0076] 在一些实施例中,样本筛选模块640包括第一样本确定模块,被配置为响应于纯度高于阈值纯度,基于簇中的样本的类别来确定经筛选样本。

[0077] 在一些实施例中,第一样本确定模块包括:第二样本确定模块,被配置为响应于簇中的样本的类别相同,将簇中的样本确定为经筛选样本。

[0078] 在一些实施例中,第一样本确定模块包括:第一数目确定模块,被配置为响应于簇中的样本具有多种类别,确定每种类别的样本数目;最大类别确定模块,被配置为基于每种类别的样本数目,确定簇中具有最大样本数目的类别;以及第三样本确定模块,被配置为将具有确定的类别的样本确定为经筛选样本。

[0079] 在一些实施例中,样本筛选模块640包括:样本比例确定模块,被配置为响应于纯度低于阈值纯度,确定簇中的样本数目占样本集中的样本总数目的比例;第二聚类模块,被配置为响应于比例超过阈值比例,对簇中的样本进行聚类,以得到聚类结果;以及第四样本确定模块,被配置为至少基于聚类结果,将簇中的样本中的至少一部分确定为经筛选样本。

[0080] 在一些实施例中,第一表示确定模块610包括:样本应用模块,被配置为将样本集应用于特征提取模型,以得到与样本集相关联的隐层神经元;以及第二表示确定模块,被配

置为基于隐层神经元,确定样本集中的样本的特征表示。

[0081] 在一些实施例中,装置600还包括:第一子集确定模块,被配置为至少基于经筛选样本,确定样本集的一个子集,子集包括从与样本集相关联的至少一个簇得到的经筛选样本;第一子集应用模块,被配置为将子集应用于特征提取模型,以得到子集中的样本的更新特征表示;第一样本更新模块,被配置为基于更新特征表示,对子集进行聚类,以基于聚类结果来更新经筛选样本。

[0082] 在一些实施例中,第一表示确定模块610包括:第三表示确定模块,被配置为确定样本集中的样本在预定义的特征空间中的特征值,作为特征表示。

[0083] 在一些实施例中,装置600还包括:第二子集确定模块,被配置为至少基于经筛选样本,确定样本集的一个子集,子集包括从与样本集相关联的至少一个簇得到的经筛选样本;以及第二样本更新模块,被配置为基于特征表示,对子集进行聚类,以基于聚类结果来更新经筛选样本。

[0084] 在一些实施例中,第一纯度确定模块630包括:第二数目确定模块,被配置为确定簇中每种类别的样本数目;最大数目确定模块,被配置为基于每种类别的样本数目,确定最大样本数目;以及第二纯度确定模块,被配置为基于最大样本数目和簇中的样本总数目确定纯度。

[0085] 图7示出了可以用来实施本公开的实施例的示例设备700的示意性框图。设备700可以用于实现图1的计算设备102。如图所示,设备700包括中央处理单元(CPU) 701,其可以根据存储在只读存储器(ROM) 702中的计算机程序指令或者从存储单元708加载到随机访问存储器(RAM) 703中的计算机程序指令,来执行各种适当的动作和处理。在RAM 703中,还可存储设备700操作所需的各种程序和数据。CPU 701、ROM 702以及RAM 703通过总线704彼此相连。输入/输出(I/O) 接口705也连接至总线704。

[0086] 设备700中的多个部件连接至I/O接口705,包括:输入单元706,例如键盘、鼠标等;输出单元707,例如各种类型的显示器、扬声器等;存储单元708,例如磁盘、光盘等;以及通信单元709,例如网卡、调制解调器、无线通信收发机等。通信单元709允许设备700通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

[0087] 处理单元701执行上文所描述的各个方法和处理,例如过程200和过程500中的任一个。例如,在一些实施例中,过程200可被实现为计算机软件程序,其被有形地包含于机器可读介质,例如存储单元708。在一些实施例中,计算机程序的部分或者全部可以经由ROM 702和/或通信单元709而被载入和/或安装到设备700上。当计算机程序加载到RAM 703并由CPU 701执行时,可以执行上文描述的过程200的一个或多个步骤。备选地,在其他实施例中,CPU 701可以通过其他任何适当的方式(例如,借助于固件)而被配置为执行过程200和过程500中的任一个。

[0088] 本文中以上描述的功能可以至少部分地由一个或多个硬件逻辑部件来执行。例如,非限制性地,可以使用的示范类型的硬件逻辑部件包括:场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、芯片上系统的系统(SOC)、负载可编程逻辑设备(CPLD)等等。

[0089] 用于实施本公开的方法的程序代码可以采用一个或多个编程语言的任何组合来编写。这些程序代码可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处

理器或控制器,使得程序代码当由处理器或控制器执行时使流程图和/或框图中所规定的功能/操作被实施。程序代码可以完全在机器上执行、部分地在机器上执行,作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

[0090] 在本公开的上下文中,机器可读介质可以是有形的介质,其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备,或者上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或快闪存储器)、光纤、便捷式紧凑盘只读存储器(CD-ROM)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

[0091] 此外,虽然采用特定次序描绘了各操作,但是这应当理解为要求这样操作以所示出的特定次序或以顺序次序执行,或者要求所有图示的操作应被执行以取得期望的结果。在一定环境下,多任务和并行处理可能是有利的。同样地,虽然在上面论述中包含了若干具体实现细节,但是这些不应当被解释为对本公开的范围的限制。在单独的实施例的上下文中描述的某些特征还可以组合地实现在单个实现中。相反地,在单个实现的上下文中描述的各种特征也可以单独地或以任何合适的子组合的方式实现在多个实现中。

[0092] 尽管已经采用特定于结构特征和/或方法逻辑动作的语言描述了本主题,但是应当理解所附权利要求书中所限定的主题未必局限于上面描述的特定特征或动作。相反,上面所描述的特定特征和动作仅仅是实现权利要求书的示例形式。

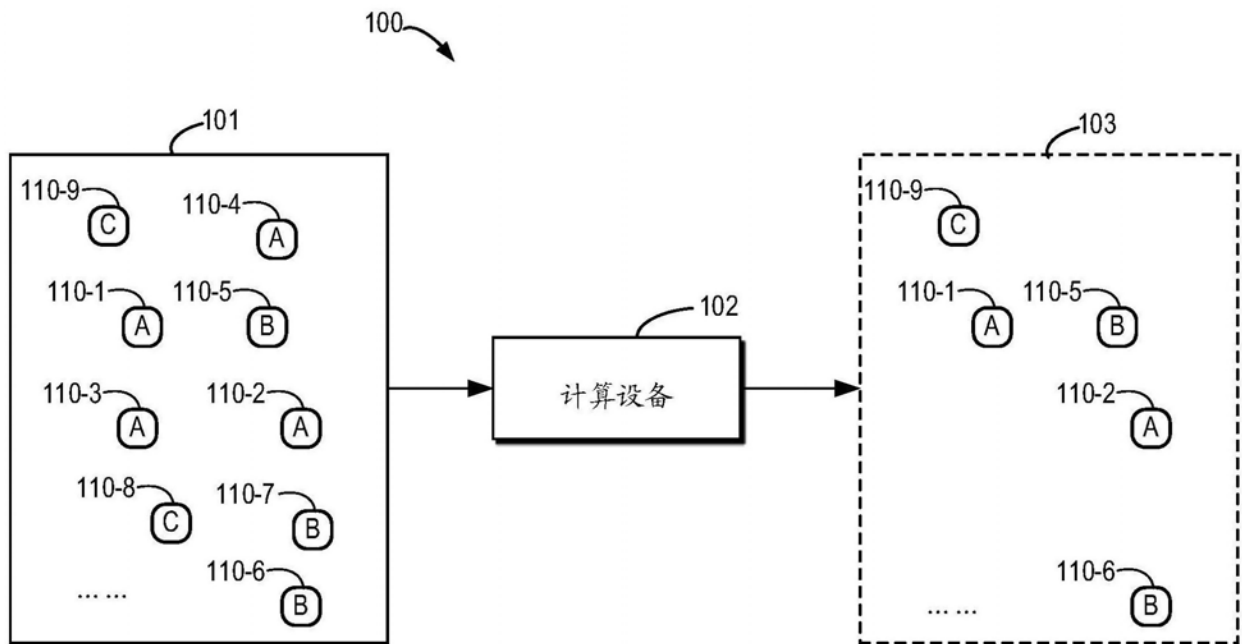


图1

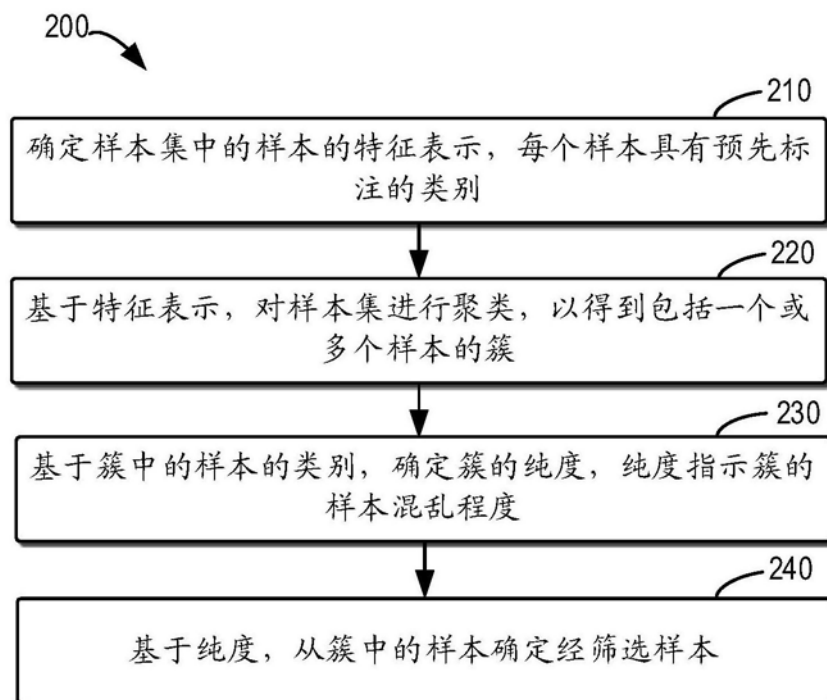


图2

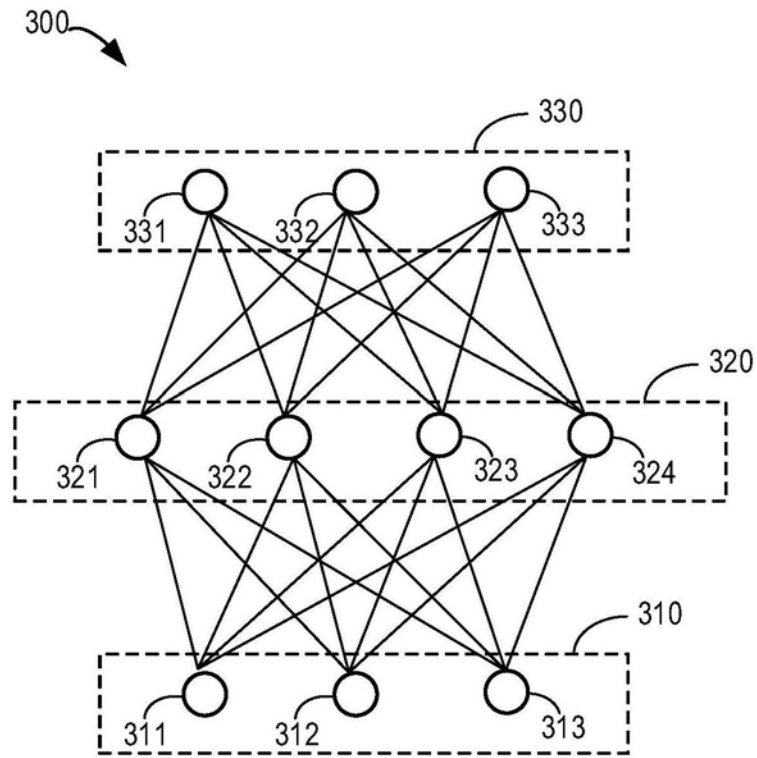


图3

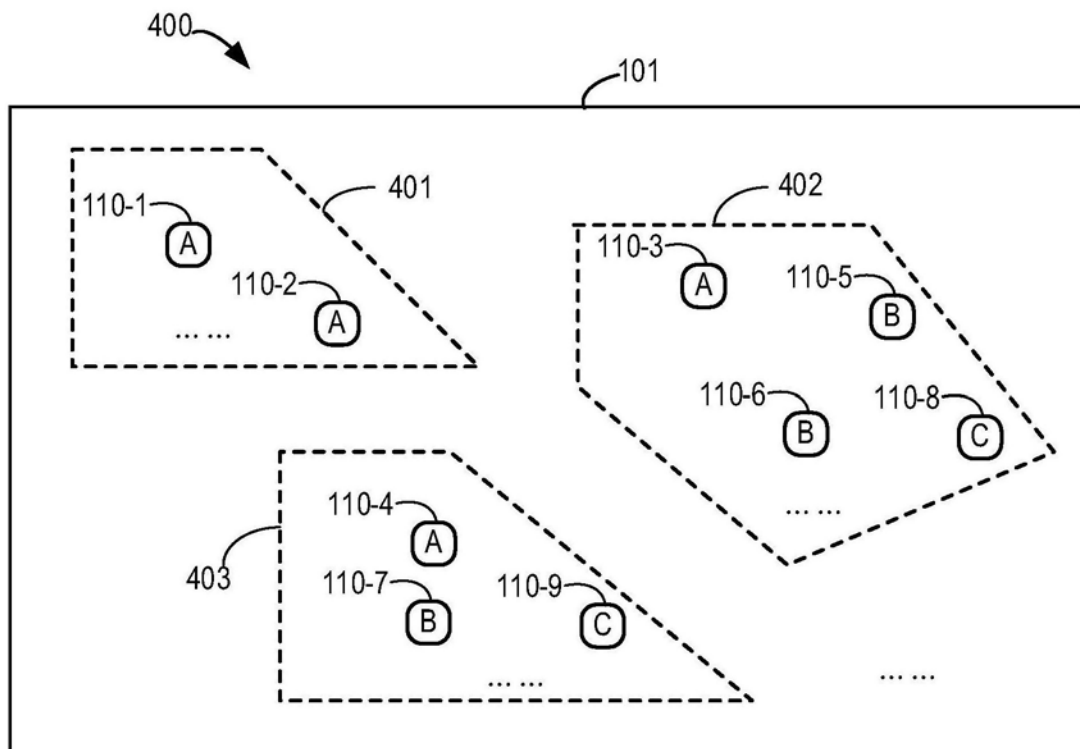


图4

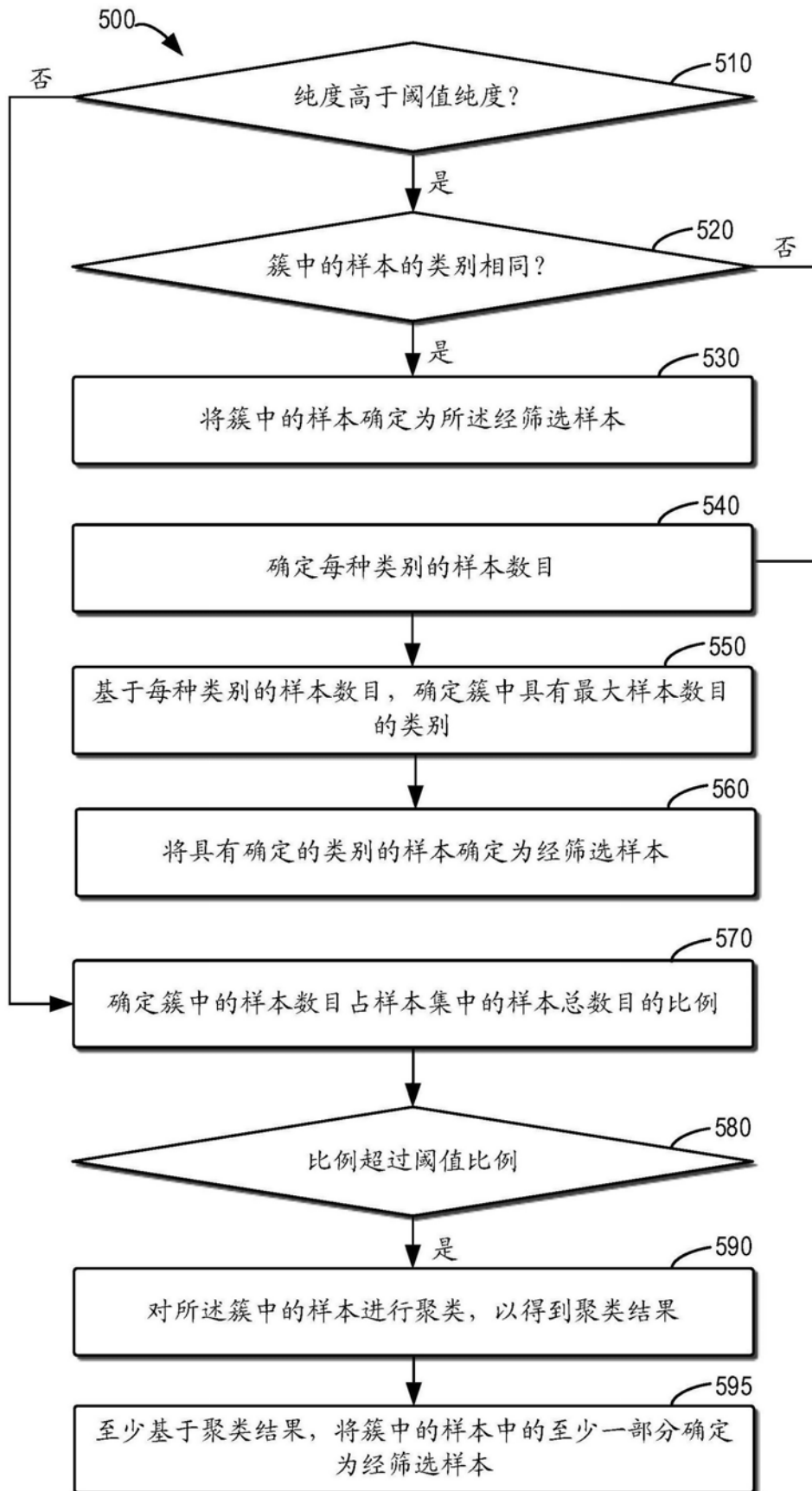


图5

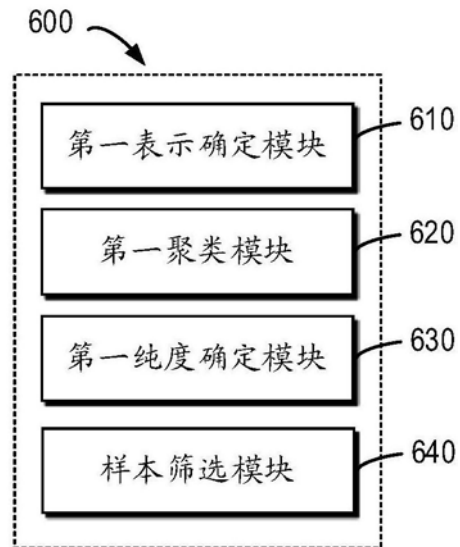


图6

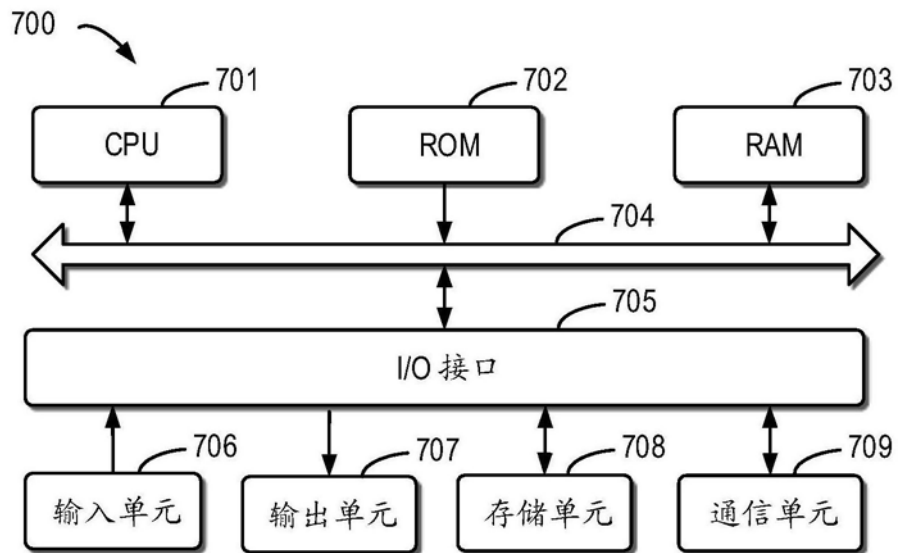


图7