



(12)发明专利申请

(10)申请公布号 CN 110249341 A

(43)申请公布日 2019.09.17

(21)申请号 201880010047.2

(74)专利代理机构 永新专利商标代理有限公司
72002

(22)申请日 2018.02.02

代理人 刘兆君

(30)优先权数据

62/454085 2017.02.03 US

(51)Int.Cl.

G06K 9/62(2006.01)

(85)PCT国际申请进入国家阶段日

2019.08.02

(86)PCT国际申请的申请数据

PCT/EP2018/052719 2018.02.02

(87)PCT国际申请的公布数据

W02018/141942 EN 2018.08.09

(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 A·卡迪尔 V·V·达特拉

K·M·Y·李 S·S·阿尔哈桑

O·F·法里

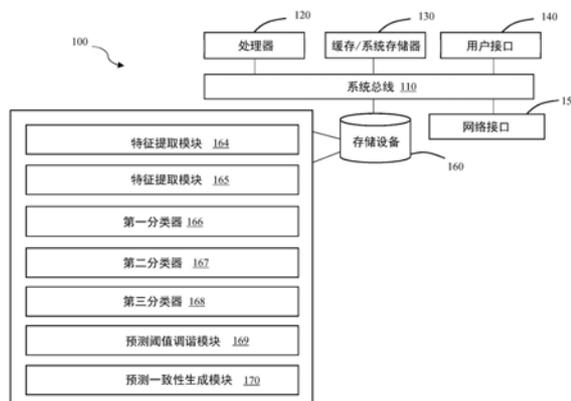
权利要求书3页 说明书12页 附图10页

(54)发明名称

分类器训练

(57)摘要

用于训练分类器的方法和系统。所述系统包括两个或更多个分类器,每个分类器可以分析从输入数据中提取的特征。所述系统可以基于所述第一标签和所述第二标签确定输入数据的真实标签,并且基于包括输入数据和真实标签的训练示例来重新训练所述第一分类器和所述第二分类器中的至少一个。



1. 一种训练分类器的方法,所述方法包括:
 - 接收标记的输入数据和未标记的输入数据;
 - 从所述标记的输入数据中提取属于第一特征空间的特征的第一集合;
 - 从所述标记的输入数据中提取属于与所述第一特征空间不同的第二特征空间的特征的第二集合;
 - 使用第一特征集训练第一分类器并将经训练的第一分类器应用于所述未标记的输入数据以预测第一标签;
 - 使用第二特征集训练第二分类器并将经训练的第二分类器应用于所述未标记的输入数据以预测第二标签;
 - 基于所述第一标签和所述第二标签来确定针对所述未标记的输入数据的真实标签;
 - 利用补充的未标记的数据以及其真实标签来扩展所述标记的输入数据;并且
 - 基于包括经扩展的标记的输入数据和所述真实标签的训练范例来重新训练所述第一分类器和所述第二分类器中的至少一个。
2. 根据权利要求1所述的方法,还包括:
 - 从所述标记的输入数据中提取属于与所述第一特征空间和所述第二特征空间不同的第三特征空间的特征的第三集合;并且
 - 使用第三特征集训练第三分类器并将经训练的第三分类器应用于所述未标记的输入数据以预测第三标签。
3. 根据权利要求2所述的方法,其中,基于所述第一标签和所述第二标签来确定针对所述未标记的输入数据的真实标签包括识别所述第一标签、所述第二标签和所述第三标签中的一致性标签。
4. 根据权利要求3所述的方法,其中,识别所述一致性标签包括:
 - 根据与所述第一分类器、所述第二分类器和所述第三分类器相关联的相应权重对所述第一标签、所述第二标签和所述第三标签中的每个进行加权,以产生针对每个唯一标签的加权投票;并且
 - 选择具有最高加权投票的唯一标签。
5. 根据权利要求4所述的方法,还包括:基于所述第一分类器、所述第二分类器和所述第三分类器相对于注释数据集的相应性能来生成针对所述第一分类器、所述第二分类器和所述第三分类器中的每个的权重。
6. 根据权利要求2所述的方法,其中,特征的所述第三集合选自包括词汇特征、语义特征和基于分布的特征的组。
7. 根据权利要求1所述的方法,其中,特征的所述第一集合和特征的所述第二集合选自包括词汇特征、语义特征和基于分布的特征的组,其中,特征的所述第一集合与特征的所述第二集合不同。
8. 一种用于训练分类器的系统,所述系统包括:
 - 接口,其用于接收标记的输入数据和未标记的输入数据;
 - 至少一个特征提取模块,其运行被存储在存储器中的指令以:
 - 从所述标记的输入数据中提取属于第一特征空间的特征的第一集合,以及
 - 从所述标记的输入数据中提取属于第二特征空间的特征的第二集合;

第一分类器,其是使用第一特征集训练的并且被配置为预测与所述未标记的输入数据相关联的第一标签;

第二分类器,其是使用第二特征集训练的并且被配置为预测与所述未标记的输入数据相关联的第二标签;以及

预测一致性生成模块,其被配置为:

基于所述第一标签和所述第二标签来确定针对所述未标记的输入数据的真实标签,并且

基于包括经扩展的输入数据和所述真实标签的训练范例来重新训练所述第一分类器和所述第二分类器中的至少一个。

9. 根据权利要求8所述的系统,其中,所述至少一个特征提取模块还被配置为从所述标记的输入数据中提取属于与所述第一特征空间和所述第二特征空间不同的第三特征空间的特征的第三集合,并且所述系统还包括第三分类器,所述第三分类器是使用第三特征集训练的并且被配置为预测与所述未标记的输入数据相关联的第三标签。

10. 根据权利要求9所述的系统,其中,所述预测一致性生成模块通过识别所述第一标签、所述第二标签和所述第三标签中的一致性标签,来基于所述第一标签和所述第二标签确定针对所述输入数据的真实标签。

11. 根据权利要求10所述的系统,其中,所述预测一致性生成模块还被配置为:

根据与所述第一分类器、所述第二分类器和所述第三分类器相关联的相应权重对所述第一标签、所述第二标签和所述第三标签中的每个进行加权,以产生针对每个唯一标签的加权投票;并且

选择具有最高加权投票的唯一标签作为所述一致性标签。

12. 根据权利要求11所述的系统,其中,所述预测一致性生成模块基于所述第一分类器、所述第二分类器和所述第三分类器相对于注释数据集的相应性能来生成针对所述第一分类器、所述第二分类器和所述第三分类器中的每个的权重。

13. 根据权利要求9所述的系统,其中,特征的所述第三集合选自包括词汇特征、语义特征和基于分布的特征的组。

14. 根据权利要求8所述的系统,其中,特征的所述第一集合和特征的所述第二集合选自包括词汇特征、语义特征和基于分布的特征的组,其中,特征的所述第一集合与特征的所述第二集合不同。

15. 一种包含用于训练分类器的计算机可执行指令的计算机可读介质,所述介质包括:

用于接收标记的输入数据和未标记的输入数据的计算机可执行指令;

用于从所述标记的输入数据中提取属于第一特征空间的特征的第一集合的计算机可执行指令;

用于从所述标记的输入数据中提取属于与所述第一特征空间不同的第二特征空间的特征的第二集合的计算机可执行指令;

用于使用第一特征集训练第一分类器并将经训练的第一分类器应用于所述未标记的输入数据以预测第一标签的计算机可执行指令;

用于使用第二特征集训练第二分类器并将经训练的第二分类器应用于所述未标记的输入数据以预测第二标签的计算机可执行指令;

用于根据所述第一标签和所述第二标签来确定针对所述未标记的输入数据的真实标签的计算机可执行指令；

用于利用补充的未标记的数据以及其真实标签来扩展所述标记的输入数据的计算机可执行指令；以及

用于基于包括经扩展的标记的输入数据和所述真实标签的训练范例来重新训练所述第一分类器和所述第二分类器中的至少一个的计算机可执行指令。

分类器训练

[0001] 相关申请的交叉引用

[0002] 本申请要求2017年2月3日提交的共同未决的美国临时申请no.62/454085的优先权和权益,其全部公开内容通过引用在此并入,如同其全部内容在本文中阐述。

技术领域

[0003] 本文中描述的实施例总体涉及用于训练分类器的系统和方法,并且更具体地但非排他地,涉及用于使用多个模型来训练分类器的系统和方法。

背景技术

[0004] 社交媒体平台提供丰富的数据来源,用于分析人们公开分享给他人的情绪。这些平台允许人们公开分享个人经历、新闻或感受,因此是丰富的信息来源,可以提供有关他们的偏好和情绪健康的宝贵见解。

[0005] 除了社交媒体之外,许多其他形式的文本和新闻报道、文章或标题的评论也可以反映和诱发情绪。可以分析这些评论和文本,以了解有新闻价值的事件如何影响人们的情绪状态和整体幸福状况。

[0006] 虽然情感极性分析已经成为研究人员的主流兴趣领域之一,但是在社交媒体条目或交互中识别更精细的情感维度(例如,快乐、愤怒、悲伤)的能力具有许多实际应用。特别是可以从对人的情绪和幸福状况的更好理解中受益的一个应用是健康护理领域。

[0007] 例如,这些知识可以帮助识别患有双相情感障碍或抑郁症的有风险的个体,有自杀倾向的个体或有愤怒管理问题的个体。此外,该知识可以帮助识别可以为这些有风险的个体触发这些条件的事件/新闻。

[0008] 为了识别情感,监督分类程序可以关于诸如情感、愤怒、恐惧、快乐、悲伤等主要情绪对来自社交媒体消息、评论、博客、新闻文章等中的文本内容进行分类。监督分类算法通常需要:(1)足够的训练数据,手动对其进行注释成本高昂;以及(2)表征/模拟问题类别的广泛的特征工程,这通常需要领域专家。

[0009] 另外,这些监督分类流程传统上没有任何用于纠错的内置机制或通过学习未标记的数据来自我改进的手段。这些技术还在单个特征空间中构建组合模型,并且因此无法利用数据集的不同独立视图。

[0010] 近年来,诸如卷积神经网络(CNN)或长短期记忆网络(LSTM)的各种深度学习模型已经在几个文本分类任务中取得了成功。但是,它们还需要大量带注释的数据集进行训练。

[0011] 半监督算法(例如,自我训练,协同训练算法)不断地识别和添加用于对模型进行训练的新的训练实例。然而,它们通常不能生成新颖或多样化的训练数据(例如,在自我训练中)。另一个缺点是错误可以通过迭代传播(例如,在共同训练中)。

[0012] 因此,需要一种克服现有技术的缺点的用于训练分类器的系统和方法。

发明内容

[0013] 提供本发明内容以用简化形式介绍一些概念,这些概念将在下面的具体实施例部分进一步描述。本发明内容并非旨在识别或排除要求保护的主题的关键特征或必要特征,也不旨在用于帮助确定要求保护的主题的范围。

[0014] 根据一个方面,各实施例涉及一种训练分类器的方法。该方法包括接收标记的输入数据和未标记的输入数据;从所述标记的输入数据中提取属于第一特征空间的特征的第一集合;从所述标记的输入数据中提取属于与所述第一特征空间不同的第二特征空间的特征的第二集合;使用第一特征集训练第一分类器并将经训练的第一分类器应用于所述未标记的输入数据以预测第一标签;使用第二特征集训练第二分类器并将经训练的第二分类器应用于所述未标记的输入数据以预测第二标签;基于所述第一标签和所述第二标签来确定未标记输入数据的真实标签;使用补充的未标记的数据及其真实标签来扩展所述标记的输入数据;并且基于包括扩展的标记输入数据和真实标签的训练示例来重新训练第一分类器和第二分类器中的至少一个。

[0015] 在一些实施例中,所述方法还包括从标记的输入数据中提取属于与所述第一特征空间和所述第二特征空间不同的第三特征空间的特征的第三集合;并且使用第三特征集训练第三分类器并将经训练的第三分类器应用于所述未标记的输入数据以预测第三标签。在一些实施例中,基于所述第一标签和所述第二标签来确定所述未标记的输入数据的真实标签包括识别所述第一标签、所述第二标签和所述第三标签中的一致性标签。在一些实施例中,识别所述一致性标签包括根据与所述第一分类器、所述第二分类器和所述第三分类器相关联的相应权重对所述第一标签、所述第二标签和所述第三标签中的每个进行加权,以产生针对每个唯一标签的加权投票;并且选择具有最高加权投票的唯一标签。在一些实施例中,所述方法还包括基于所述第一分类器、所述第二分类器和所述第三分类器相对于注释数据集的相应性能,生成针对所述第一分类器、所述第二分类器和所述第三分类器中的每个的权重。

[0016] 在一些实施例中,特征的所述第三集合选自包括词汇特征、语义特征和基于分布的特征的组。

[0017] 在一些实施例中,特征的第一集合和特征的第二集合选自包括词汇特征、语义特征和基于分布的特征的组,其中,特征的第一集合与特征的第二集合不同。

[0018] 根据另一方面,各实施例涉及一种用于训练分类器的系统。所述系统包括:用于接收标记的输入数据和未标记的输入数据的接口;至少一个特征提取模块,其执行存储在存储器中的指令,以从所述标记的输入数据中提取属于第一特征空间的特征的第一集合,并且从所述标记的输入数据中提取属于第二特征空间的特征的第二集合;第一分类器,其是使用所述第一特征集训练的并且被配置为预测与所述未标记的输入数据相关联的第一标签;第二分类器,其是使用第二特征集训练的并且被配置为预测与所述未标记的输入数据相关联的第二标签;以及预测一致性生成模块,其被配置为基于所述第一标签和所述第二标签确定未标记输入数据的真实标签,并且基于包括扩展输入数据和真实标签的训练范例来重新训练所述第一分类器和所述第二分类器中的至少一个。

[0019] 在一些实施例中,所述至少一个特征提取模块还用于提取属于与所述第一特征空间和所述第二特征空间不同的第三特征空间的特征的第三集合,并且所述系统还包括第三

分类器,所述第三分类器被配置为输出与特征的第三集合相关联的第三标签。在一些实施例中,预测一致性生成模块通过识别第一标签、第二标签和第三标签中的一致性标签,来基于第一标签和第二标签确定针对所述输入数据的真实标签。在一些实施例中,所述预测一致性生成模块还被配置为根据与所述第一分类器、所述第二分类器和所述第三分类器相关联的相应权重对所述第一标签、所述第二标签和所述第三标签中的每个进行加权,以产生针对每个唯一标签的加权投票;并且选择具有最高加权投票的唯一标签作为一致性标签。在一些实施例中,所述预测一致性生成模块基于所述第一分类器、所述第二分类器和所述第三分类器相对于注释数据集的相应性能,来针对所述第一分类器、所述第二分类器和所述第三分类器中的每个生成权重。在一些实施例中,特征的所述第三集合选自包括词汇特征、语义特征和基于分布的特征的组。

[0020] 在一些实施例中,特征的第一集合和特征的第二集合选自包括词汇特征、语义特征和基于分布的特征的组,其中,特征的第一集合与特征的第二集合不同。

[0021] 根据又一方面,各实施例涉及一种包含用于训练分类器的计算机可执行指令的计算机可读介质。所述介质包括:用于接收输入数据的计算机可执行指令;用于从输入数据中提取属于第一特征空间的特征的第一集合的计算机可执行指令;从的输入数据中提取属于不同于所述第一特征空间的第二特征空间的特征的第二集合的计算机可执行指令;用于将第一分类器应用于第一特征集以接收第一标签的计算机可执行指令;用于将第二分类器应用于第二特征集以接收第二标签的计算机可执行指令;用于根据第一标签和第二标签来确定针对所述输入数据的真实标签的计算机可执行指令;以及用于基于包括输入数据和真实标签的训练范例来重新训练所述第一分类器和所述第二分类器中的至少一个的计算机可执行指令。

附图说明

[0022] 参考以下附图描述本文中的实施例的非限制性和非穷举性实施例,其中,除非另有说明,否则相同的附图标记在各个视图中指代相同的部件。

[0023] 图1图示了根据一个实施例的用于训练分类器的系统;

[0024] 图2图示了根据一个实施例的图1的部件的工作流程;

[0025] 图3图示了根据一个实施例的图1的第一分类器的工作流程;

[0026] 图4图示了根据一个实施例的图1的第二分类器的工作流程;

[0027] 图5图示了根据一个实施例的图1的第三分类器的工作流程;

[0028] 图6图示了根据一个实施例的图1的预测阈值调谐模块的工作流程;

[0029] 图7图示了根据一个实施例的图1的预测一致性生成模块的工作流程;并且

[0030] 图8描绘了根据一个实施例的用于训练分类器的方法的流程图;

[0031] 图9图示了根据另一实施例的用于训练分类器的系统;并且

[0032] 图10描绘了根据一个实施例的使用图9的系统训练分类器的方法的流程图。

具体实施方式

[0033] 以下参考附图更详细地描述了各种实施例,附图形成实施例的一部分,并且示出了具体的示例性实施例。然而,本公开的概念可以以许多不同的形式实现,并且不应该被解

释为限于本文中阐述的实施例；相反，这些实施例是作为全面和完整公开的一部分提供的，以向本领域技术人员充分传达本公开的概念、技术和实现的范围。各实施例可以被实践为方法、系统或设备。因此，实施例可以采取硬件实现方式、完全软件实现方式或组合了软件和硬件方面的实现方式的形式。因此，以下详细描述不应被认为是限制性的。

[0034] 说明书中对“一个实施例”或“实施例”的引用意味着结合所述实施例描述的特定特征、结构或特性被包括在根据本公开的至少一个范例实现方式中。在说明书中的各位置，短语“在一个实施例中”的出现不一定都指代相同的实施例。在说明书中的各位置，短语“在一些实施例中”的出现不一定都指代相同的实施例。

[0035] 以下说明的一些部关于存储在计算机存储器中的非瞬态信号的操作的符号表示来呈现。这些描述和表示是数据处理领域的技术人员用来最有效地将其工作的本质传达给本领域其他技术人员。这样的操作通常需要对物理量的物理操作。通常，但是不一定，这些量采取能够被存储、传输、组合、比较和以其他方式操纵的电、磁或光学信号的形式。有时，主要是出于通常使用的原因，将这些信号称为比特、值、元素、符号、字符、术语、数字等。此外，有时方便起见，也将需要对物理量的物理操纵的步骤的特定布置称为模块或代码设备，而不失一般性。

[0036] 然而，所有这些和类似的术语都将与适当的物理量相关联，并且仅仅是应用于这些量的便利标签。除非根据以下讨论显而易见另行专门陈述，应该理解，在贯穿本说明，使用诸如“处理”或“运算”或“计算”或“确定”或“显示”等术语的讨论是指计算机系统或类似电子计算设备的动作和过程，其对表示为计算机系统存储器或寄存器或其他此类信息存储、传输或显示设备内的物理（电子）量的数据进行操纵和变换。本公开的部分包括可以以软件、固件或硬件实现的过程和指令，并且当以软件实现时，可以将其下载以驻留在各种操作系统所使用的不同平台上并由其操作。

[0037] 本公开还涉及一种用于执行本文中的操作的装置。该装置可以针对期望的目的而特别构造，或者其可以包括由存储在计算机中的计算机程序选择性地激活或重新配置的通用计算机。这样的计算机程序可以存储在计算机可读存储介质中，诸如但不限于任何类型的盘、包括软盘、光盘、CD-ROM、磁光盘、只读存储器（ROM）随机存取存储器（RAM）、EPROM、EEPROM、磁卡或光卡、专用集成电路（ASIC）、或适用于存储电子指令的任何类型的介质，并且每个都可以被耦合到计算机系统总线。此外，说明书中所指的计算机可以包括单个处理器，或者可以是采用多个处理器设计以提高计算能力的架构。

[0038] 本文提出的过程和显示并不固有地与任何特定的计算机或其他装置相关。各种通用系统也可以根据本文的教导的程序一起使用，或者构造更专用的装置来执行一个或多个方法步骤可以证明是方便的。在下面的描述中讨论了针对各种这样的系统的结构。另外，可以使用足以实现本公开的技术和实现方式的任何特定编程语言。可以使用各种编程语言来实现如本文中所讨论的本公开。

[0039] 此外，说明书中使用的语言主要是为了可读性和指导性目的而选择的，并且可能没有被选择来描绘或限定本公开的主题。因此，本公开旨在说明而非限制本文中所讨论的概念的范围。

[0040] 本文中描述的实施例提供了迭代框架，其可以将分类器与特征空间的不同视图组合。在一些实施例中，例如用于基于社交媒体内容对情绪进行分类的实施例，这些分类器可

以包括：(1) 基于词汇特征的分类器；(2) 基于语义特征的分类器；以及(3) 基于分布特征的分类器。然后，这些分类器可以对分类标签进行投票，然后可以使用该分类标签在将来的迭代中进一步训练分类器。

[0041] 这种基于系集(ensemble)的框架提供了两个主要的好处。首先，由于与另一个分类器的一致性，这些实施例为任何分类器提供了纠错机会。例如，如果第一分类器错误地预测针对推文(tweet)的情绪e，但是第二和/或第三分类器没有，则不将推文并入到用于下一次迭代的训练数据中，因此避免了可能通过相继的迭代传播的潜在错误。这与现有的共同训练技术形成对比，在现有的共同训练技术中，仍将提供推文作为第二和第三分类器的训练实例。

[0042] 第二个优点是分类器可以获得分类器可能无法自己识别的新训练实例。例如，如果第一分类器未能预测针对推文的情绪e，并且第二和第三分类器预测了针对推文的e，则仍将提供推文作为下一次迭代的第一分类器的训练实例。这与传统的自我训练技术不同，在传统的自我训练技术中，如果分类器不识别推文的情绪e，则推文不会被添加到训练集中以用于下一次迭代。

[0043] 图1图示了根据一个实施例的用于训练分类器的系统100。系统100可以包括经由一个或多个系统总线110互连的处理器120、存储器130、用户接口140、网络接口150和存储设备160。应该理解，在一些方面中，图1构成抽象，并且系统100的实际组织及其部件可以与所示的不同。

[0044] 处理器120可以是能够执行存储在存储器130、存储设备160上或者是能够处理数据的指令的任何硬件设备。这样，处理器120可以包括微处理器、现场可编程门阵列(FPGA)、专用集成电路(ASIC)或其他类似设备。

[0045] 存储器130可以包括各种非瞬态存储器，例如L1、L2或L3高速缓存或系统存储器。这样，存储器130可以包括静态随机存取存储器(SRAM)、动态RAM(DRAM)、闪存、只读存储器(ROM)或其他类似的存储器设备和配置。只要可以执行用于训练(一个或多个)分类器的指令，存储器130的确切配置就可以改变。

[0046] 用户接口140可以包括用于实现与用户的通信的一个或多个设备。例如，用户接口140可以包括用于接收用户命令的显示器、鼠标和键盘。在一些实施例中，用户接口140可以包括命令行接口或图形用户接口，其可以经由网络接口150呈现给远程终端。用户接口140可以在诸如PC、膝上型电脑、平板电脑、移动设备等的用户设备上执行。

[0047] 网络接口150可以包括用于实现与其他远程设备的通信的一个或多个设备。例如，网络接口150可以包括被配置为根据以太网协议进行通信的网络接口卡(NIC)。另外，网络接口150可以根据TCP/IP协议实现用于通信的TCP/IP栈。用于网络接口150的各种替代或附加硬件或配置将是显而易见的。网络接口150可以与诸如社交媒体平台的各种源连接或以其他方式接收数据。

[0048] 存储设备160可以包括一个或多个机器可读存储介质，诸如只读存储器(ROM)，随机存取存储器(RAM)，磁盘存储介质，光存储介质，闪存设备或类似的存储介质。在各种实施例中，存储设备160可以存储用于由处理器120执行的指令或模块或者处理器120可以操作的数据。

[0049] 例如，存储设备160可以包括一个或多个特征提取模块164和165，第一分类器166、第

二分类器167、第三分类器168、预测阈值调整模块169和预测一致性生成模块170。作为存储器160的一部分而被包括的确切部件可以变化,并且可以包括除图1中所示的那些之外或代替其的其他部件。额外地或替代地,单个部件可以执行图1中所示的多于一个部件的功能。

[0050] 特征提取模块164和165可以从数据集中提取特定特征以供分类器分析。尽管图1中示出了两个特征提取模块,但是特征提取模块的数量可以变化。例如,可以存在与每个分类器相关联的一个特征提取模块。或者,单个特征提取模块可以被配置为提取针对每个分类器的特定特征。特征提取模块164在本申请的其余部分中将被描述为执行特征提取功能。

[0051] 在用于对情绪进行分类的实施例中,第一分类器166可以是基于词汇特征的分类器。例如,第一分类器166可以对所接收的数据集使用词袋建模过程。

[0052] 第二分类器167可以考虑社交媒体条目的基于语义的特征。为了对语义特征空间进行建模,第二分类器167可以使用来自表示语义空间中的专家知识的知识库的语义关系,以及为了利用表示语义关系的分布相似性度量而创建的语义关系。

[0053] 第二分类器167可以对出现在合适的知识库(例如,WORDNET)中的任何单词/术语使用二元特征,以及与社交媒体条目中的单词的上位词、下位词、局部词、动词组或与“类似”的关系。这些关系中的每一个可以表示唯一的特征类型。

[0054] 例如,“汽车”与“机动车辆”具有上位关系,与“窗”具有局部关系。如果“汽车”在社交媒体条目中显示为单词,则二元特征可以表示关系-术语对“上位:机动车辆”,另一个二元特征可能代表关系-术语对“局部:窗。”社交媒体条目中使用的单词意义没有消除歧义,而是所有意义都可以用作语义特征字典的一部分。

[0055] 另外,可以使用语义上相似的单词(通过分布相似性度量确定为额外语义特征)。在大的数据集上训练的单词嵌入模型可以用于针对社交媒体条目中的每个单词,使用与嵌入向量的余弦相似性来检索二十(20)个最相似的单词。然后,可以检索针对社交媒体条目中的单词的每个语义上相似的单词的二元特征。

[0056] 第三分类器168可以是基于分布特征的分类器。例如,第三分类器168可以使用现有的情绪和情感词典,并且考虑具有种子情感令牌的推文中的单词的分布相似性。

[0057] 为了生成分布特征的第一集合,第三分类器168可以使用本领域中已知的情绪指示符词典。词典可以包含情感主题标签,主题标签模式以及根据主题标签和模式创建的情感短语。指示符可能属于五种情绪类别之一:(1)情感;(2)生气/愤怒;(3)恐惧/焦虑;(4)快乐;以及(5)悲伤/失望。针对情绪的每个指示符,第三分类器168可以创建一个二元特征。针对给定的推文或社交媒体条目,如果推文包含来自相应情绪词典之一的短语或主题标签,则将特征值设置为“1”。

[0058] 在一些实施例中,可以使用两个单词情感词典的集合,其考虑使用众包创建的词典和使用自动方法创建的词典。词典可以包含关于各种情绪(例如,愤怒,恐惧,期待,信任,惊讶,悲伤,喜悦,以及厌恶)和两种情绪(负面或正面)的单词关联(例如,二元或真实值分数)。对于给定的社交媒体条目,如果条目包含来自与上述八种情绪之一相关联的词典之一的单词,则可以将特征值设置为1。

[0059] 在一些实施例中,另一组分布特征可以使用AFINN情感词典,其包含具有正的或负的情绪分数的2477个单词。利用词典,第三分类器168可以使用两个二元特征,一个针对正的,一个针对负的。对于给定的社交媒体条目,如果条目包含在AFINN词典中具有正值或负

值的单词,则将特征值设置为1。

[0060] 在一些实施例中,第三分类器168可以确定社交媒体条目中的单词与种子情感令牌分布相似性。为了对具有情感类别的条目的分布相似性进行建模,第三分类器168可以使用情感类别的种子令牌并且确定它们与分布空间中的条目的词的余弦相似性。

[0061] S可以是一组有序的种子情感令牌并且T可以是推文中的一组单词。第三分类器168可以通过以下方式创建向量作为关于前面提到的情感类别的推文的分布表示:

[0062] $\text{Dist}(\text{seeds}_s, \text{tweet}) = \text{argmax}_{\text{Cosine}(\text{seeds}_s, x)}, x \in T$

[0063] 在这种情况下, sinS 是注释类别的种子令牌,并且 $\text{Dist}(\text{seeds}_s, \text{tweet})$ 函数代表向量的第s个元素。

[0064] 图2图示了根据一个实施例的图1的部件的工作流程200。在该实施例中,带注释(即标记)的训练数据202可以包括推文、博客、新闻文章、标题等。同样,在基于社交媒体内容对情绪进行分类的背景下描述了该实施例。然而,所述体系结构可以扩展到在其他类型的应用程序或域中训练分类器。

[0065] 分类器166、167和168可以接收用于监督训练的注释训练数据202。如前所述,第一分类器166可以是基于词汇特征的分类器、第二分类器167可以是基于语义特征的分类器,第三分类器168可以是基于分布特征的分类器。在对带注释的训练数据202进行监督训练之后,分类器166、167和168可以各自提供训练的分类模型。

[0066] 经训练的分类器166、167和168的分类模型可以在专家注释的训练数据204上执行,以供预测阈值调整模块169进一步改进。预测阈值调整模块169可将每个分类器模型应用于提出的专家注释调谐数据204以确定高置信度预测阈值。

[0067] 然后,训练的分类器166、167和168的分类模型可以分析未标记的数据206以进行分类。该未标记的数据206可以包括大量社交媒体条目、推文、博客、新闻文章、标题等。每个分类器166、167和168可以输出指示他们是否认为社交媒体条目与情绪e相关联的标签。

[0068] 预测一致性生成模块170可以对来自分类器166、167和168的分类决策进行加权投票或多数投票,并输出关于未标记的数据206的预测。预测一致性生成模块170的输出可以包含在训练数据202中并且重复该过程。因此,注释数据集202的大小随着每次迭代而增加,并且未标记的数据集206的大小随着每次迭代而减小。可以重复该过程直到满足停止准则。

[0069] 图2的架构200包括可以适于添加更多分类器作为使用不同分类过程的整体的一部分。例如,具有特征工程的支持向量机(SVM),逻辑回归(LR)等,或者诸如没有特征设计等的卷积神经网络(CNN)的神经网络分类模型可以用于实现本文描述的各种实施例的特征。

[0070] 图3示出了根据一个实施例的第一分类器166的工作流程300。如前所述,第一分类器166可以考虑数据集202的词汇视图。数据集202可以被提供给特征提取模块164,并且可以是包括社交媒体条目的注释训练数据集,所述社交媒体条目包括推文,博客,评论,新闻文章,标题等,以及关于用户对这样的反应的数据。然后,特征提取模块164可以从数据集202中提取词袋特征,其可以被传送到第一分类器166以用于监督学习。

[0071] 作为使用词袋特征的监督学习过程的结果,第一分类器166可以执行第一训练分类模型304。模型304可以基于例如逻辑回归分析来考虑分配给某些特征的特定权重。这些权重实质上告诉系统特定功能的重要性。然后,第一分类器166的训练分类模型304可以作

为调谐过程的一部分在专家注释数据204上执行,以及在未标记的数据206上执行以输出预测概率308。

[0072] 图4图示了根据一个实施例的第二分类器167的工作流程400。如前所述,第二分类器167可以考虑数据集202的语义视图(其可以是图3的相同数据集202)。

[0073] 特征提取模块164可以从预先训练的词嵌入的一个或多个数据库404接收从分布矢量空间确定的语义上相似的词。第二分类器167还可以接收关于数据集202中的单词的语义关系的数据(例如,上位词,局部词,全称,下位词,动词组,类似词,同义词,反义词等)。可以从一个或多个语义知识数据库406(例如WordNet)检索关于语义关系的这种类型的数据。

[0074] 可以将提取的语义特征传送到第二分类器167以进行监督学习。作为监督学习过程的结果,第二分类器167可以执行第二训练分类模型408。训练的分类模型408可以基于例如逻辑回归分析来考虑分配给某些特征的特定权重。这些权重实质上告诉系统特定功能的重要性。然后,第二分类器167的训练分类模型408可以作为调谐过程的一部分在专家注释数据204上执行,以及在未标记的数据206上执行以输出预测概率410。

[0075] 图5示出了根据一个实施例的第三分类器168的工作流程500。如前所述,第三分类器168可以考虑数据集202的分布特征(其可以与图3和4的数据集202相同)。

[0076] 特征提取模块164可以从数据集202提取分布特征。特征提取模块164可以从一个或多个种子词数据库504接收种子情绪词。特征提取模块164还可以从一个或多个先前训练的词嵌入数据库506接收与情绪种子词类似的词。

[0077] 特征提取模块164可以提取与(一个或多个)种子情绪词和数据集202的文本中最相似的词之间的矢量差异相关的分布特征。可以将提取的特征传送到第三分类器168以进行监督学习。

[0078] 作为监督学习过程的结果,第三分类器168因此可以执行第三训练分类模型508。训练的分类模型508可以基于例如逻辑回归分析来考虑分配给某些特征的特定权重。这些权重实质上告诉系统特定功能的重要性。然后,训练分类模型408可以作为调谐过程的一部分在专家注释数据204上执行,以及在未标记的数据206上执行以输出预测概率510。

[0079] 图6描绘了根据一个实施例的预测阈值调整模块169的工作流程600。预测阈值调整模块169可以分别从分类模型304、408和508接收与输入数据202相关联的预测概率308、410、510。

[0080] 预测阈值调整模块169可以基于其置信度得分来过滤掉或者选择特定预测。例如,预测阈值调整模块169可以选择具有前25%最高置信度值的那些预测。预测阈值调整模块169的输出可以是一组调谐的预测阈值602以确保高精度(例如,每情绪,每分类器)。

[0081] 在本申请的上下文中,“阈值”可以被定义为截止概率,在该截止概率之上,实例被分类为情绪类别。如果预测概率低于阈值,则实例不归类于情绪。

[0082] 图7示出了根据一个实施例的预测一致性生成模块170的工作流程700。分类器166、167、168的训练模型304、408和508可以分别分析未标记的数据206。未标记的数据206可以包括推文,博客,新闻文章,标题等。

[0083] 训练模型304、408和508还可以考虑由预测阈值调整模块169提供的调谐阈值702。然后,模型304、408和508可以提供分类预测,所述分类预测被传送到预测一致性生成模块170以进行加权投票过程。

[0084] 可以从带注释的验证数据204确定每个分类器166、167和168的权重。因此,预测一致性生成模块169的输出可以是高置信度注释数据702。然后可以将该高置信度注释数据702添加到带注释的训练数据202,以进一步训练分类器。因此,尺寸注释训练数据集202可以随着每次迭代而连续增加。

[0085] 图8描绘了根据一个实施例的用于训练分类器的方法800的流程图。步骤802涉及接收标记的输入数据和未标记的数据。该数据可以包括带注释的社交媒体数据,例如用户做出的推文或在线评论。

[0086] 步骤804涉及从标记的输入数据中提取属于第一特征空间的特征的第一集合。例如,步骤804可以由诸如图1的特征提取模块164的特征提取模块执行。该特征的第一集合可以包括语义特征、词典特征或分布特征。

[0087] 步骤806涉及从标记的输入数据中提取属于不同于所述第一特征空间的第二特征空间的特征的第二集合。例如,该步骤可以由诸如图1的特征提取模块164的特征提取模块执行。这些功能可能包括语义功能、词典功能或分布功能。无论提取的功能如何,特征的第二集合都应与其特征的第一集合不同。

[0088] 尽管未在图8中示出,但是一些实施例可以进一步提取属于与第一特征空间和第二特征空间不同的第三特征空间的特征的第三集合。例如,该步骤可以由诸如图1的特征提取模块164的特征提取模块执行。该特征的第三集合可以包括语义特征、词典特征或分布特征。无论提取的特征如何,特征的第三集合都应与其特征的第一集合和特征的第二集合不同。

[0089] 步骤808涉及使用第一特征集训练第一分类器并将经训练的第一分类器应用于所述未标记的输入数据以预测第一标签。第一分类器可以类似于图1的第一分类器164,例如,并且可以是基于词汇特征的分类器。第一标签可以基于第一分类器的分析来指示输入数据是否与特定情绪相关联,例如喜悦或愤怒。

[0090] 步骤810涉及使用第二特征集训练第二分类器并将经训练的第二分类器应用于所述未标记的输入数据以预测第二标签。第二分类器可以类似于图1的第二分类器167,例如,并且可以是基于语义特征的分类器。第二标签可以基于第二分类器的分析来指示输入数据是否与特定情绪相关联。

[0091] 尽管未在图8中示出,但是一些实施例还可包括使用提取的第三特征集训练第三分类器以预测第三标签的步骤。该第三分类器可以类似于图1的第三分类器168,例如,并且可以是基于分布特征的分类器。第三标签可以基于第三分类器的分析来指示输入数据是否与特定情绪相关联。

[0092] 步骤812涉及至少基于第一标签和第二标签来确定针对未标记的输入数据的真实标签。该真实标签可以是来自每个分类器的投票的结果,所述投票是关于数据是否表现出训练分类器的特定情绪。

[0093] 在一些实施例中,基于第一标签和第二标签确定输入数据的真实标签包括识别所述第一标签、所述第二标签和所述第三标签中的一致性标签。在一些实施例中,识别所述一致性标签可以涉及根据与所述第一分类器、所述第二分类器和所述第三分类器相关联的相应权重对所述第一标签、所述第二标签和所述第三标签中的每个进行加权,以产生针对每个唯一标签的加权投票。这些权重可以基于分类器相对于标记输入数据的相应性能。然后,可以选择具有最高加权投票的唯一标签作为一致性标签。

[0094] 步骤814涉及使用补充的未标记的数据及其真实标签来扩展标记的输入数据。由于此数据现在已标记,因此可将其添加到训练数据集中并用于将来的迭代。

[0095] 步骤816涉及基于包括扩展的标记输入数据和真实标签的训练示例来重新训练第一分类器和第二分类器中的至少一个。现在可以将输入的数据(其现在与真实标签相关联)添加回带注释的训练数据集。然后可以多次迭代方法800(即,添加到带注释的训练集和再训练),直到不能将新的训练示例添加到带注释的集合。

[0096] 图9示出了根据另一实施例的用于训练分类器的系统900。在该实施例中,利用特征空间的三个视图中的每个(如图1中)来独立地训练分类器以预测情绪。

[0097] 在经典的共同训练中,来自每个分类器识别的未标记的数据中最有置信度地标记的实例作为补充训练实例给予其他分类器。然而,有可能并非所有分类器都适合于将实例的正确的集合识别为其他分类器的补充数据。

[0098] 然而,图9的系统可以将三个分类器中最弱的分类识别为要改进的目标视图分类器。为了实现这一点,剩余的特征空间视图可以基于该补充视图分类器将比弱分类器执行得更好的假设来训练补充视图分类器。然后,补充视图分类器可以引导目标视图分类器利用可能被目标视图分类器错误分类的新训练数据来改进自身。

[0099] 部件910、920、930、940和950分别类似于图1的部件110、120、130、140和150,并且这里不再重复。提取模块964、965和分类器966-968分别类似于图1的部件164、165和166-168,并且这里不再重复。

[0100] 图9的系统900还可以包括视图选择模块969。视图选择模块969可以被配置为评估个体视图分类器在验证数据集上的性能,并将最弱性能的分类器指定为目标视图分类器。视图选择模块969还可以组合剩余视图(来自(一个或多个)其他分类器)以创建补充视图分类器。

[0101] 图9的系统900还可以包括实例排名模块970。实例排名模块970可以被配置为评估和组合目标视图和补充视图分类器的预测概率,以选择用于重新训练分类器的补充训练数据。

[0102] 图10描绘了根据另一个实施例的用于训练图9的多个分类器的迭代框架1000。在该特定实施例中,框架1000可用于基于社交媒体内容对用户的情绪进行分类。

[0103] 首先,在事件1002中,使用用于针对每个情感 $e \in \dot{E}$ 训练二元分类器的初始集合的接口来接收与情绪类别E(例如,情感、快乐、愤怒)相关联的先前注释的数据集(例如,诸如推文、评论、帖子等的社交媒体条目的数据集)。

[0104] 可以针对情绪e训练每个分类器966、967和968。如前所述,第一分类器966可以具有词汇视图(LEX_c)、第二分类器967可以具有语义视图(SEM_c),并且第三分类器968可以具有特征空间的分布视图(EMO_c)。

[0105] 在事件1004中,对于情绪e,分类器966、967和968可以独立地应用于先前注释的验证数据集以评估它们的表现。在事件1006中,视图选择模块969选择最弱的分类器作为具有目标视图的目标分类器。该目标分类器是为改进而选择的分类器。

[0106] 在事件1008中,视图选择模块969选择(一个或多个)其他分类器作为补充视图分类器,并用于生成目标视图的至少一个补充视图。可以仅使用其他“非目标”视图中的一个,或者可以使用和组合两个其他非目标视图以提供至少一个补充视图。目标和补充分类器都

应用于事件1010中的未标记的数据集,并且目标视图分类器和补充视图分类器可以各自为每个社交媒体条目(例如,推文)分配分类概率。

[0107] $P_t(\text{tweet})$ 可以是由目标分类器分配的概率,并且 $P_c(\text{tweet})$ 可以是由补充分类器分配的概率。为了使用这两个概率对未标记的数据进行排名,实例排名模块970可以通过执行以下功能来为特定推文分配分数:

[0108] 分数(tweet) = $P_c(\text{tweet}) \times (1 - P_t(\text{tweet}))$

[0109] 上述函数更强烈地奖励补充分类器分配高概率但是目标分类器不分配高概率的推文。这反映了目标分类器的改进机会。

[0110] 实例排名模块970可以使用由上述评分函数生成的分数对所有未标记的数据进行排序。然后,预测一致性生成模块可以选择例如原始训练数据大小的前25%(使得新数据不会压倒先前的训练数据)。在扩展原始训练数据集之后,可以重新训练分类器并重复该过程。

[0111] 具有补充视图的分类器可能已经比目标视图更好地识别验证数据集实例。因此,期望通过组合它们的特征空间,新分类器将能够比目标视图分类器更好地识别新实例。

[0112] 在事件1010中,为每个未标记的社交媒体条目(例如,推文)生成至少两个分类器输出——一个来自目标分类器,一个来自补充分类器。使用其分配的分类概率到社交媒体条目,实例排名模块970可以执行排名功能以识别目标分类器置信度较低的实例。

[0113] 然后将高度排名的社交媒体条目添加到针对特定情绪e的目标分类器的训练数据中。然后可以迭代图9中所图示的过程,直到例如满足停止准则。

[0114] 图9和图10的系统900和方法1000分别提供了两个重要的好处。首先,他们通过使用性能更好的分类器提供纠错机会。其次,它们为目标视图分类器提供了采集目标视图分类器无法使用其自己的特征空间自行识别的新训练实例的机会。

[0115] 上面讨论的方法、系统和设备是示例。各种配置可以适当地省略、替换或添加各种过程或部件。例如,在替代配置中,可以以与所描述的顺序不同的顺序执行方法,并且可以添加、省略或组合各种步骤。而且,关于某些配置描述的特征可以以各种其他配置组合。可以以类似的方式组合配置的不同方面和元素。此外,技术在发展,因此,许多元素是示例,并且不限制本公开或权利要求的范围。

[0116] 上面参照根据本公开的实施例的方法、系统和计算机程序产品的框图和/或操作说明来描述了本公开的实施例。方框中记载的功能/动作可能按照不同于任何流程图所示的顺序发生。例如,取决于所涉及的功能/动作,相继地示出的两个块实际上可以基本上同时执行,或者块有时可以以相反的顺序执行。额外地或者替代地,并不是任何流程图中示出的所有块都需要执行和/或运行。例如,如果给定的流程图具有包含功能/动作的五个块,则可能的情况是,仅执行和/或运行五个块中的三个块。在该范例中,可以执行和/或运行五个块中的任意三个。

[0117] 值超过(或大于)第一阈值的陈述等同于该值满足或超过略大于第一阈值的第二阈值的陈述,例如,第二阈值是一个值高于相关系统分辨率中的第一个阈值。值小于第一阈值(或在其之内)的陈述等同于该值小于或等于略低于第一阈值的第二阈值的陈述,例如,第二阈值是相关系统分辨率中低于第一阈值的一个值。

[0118] 在说明书中提出了具体细节以提供对示例配置(包括实现方式)的透彻理解。然

而,可以在没有这些具体细节的情况下实践配置。例如,已经示出了众所周知的电路、过程、算法、结构和技术而没有不必要的细节,以避免使配置模糊。本说明仅提供示例配置,并且不限制权利要求的范围、适用性或配置。相反,前面对配置的描述将为本领域技术人员提供用于实现所描述的技术的实现的描述。在不脱离本公开的精神或范围的情况下,可以对元件的功能和布置进行各种改变。

[0119] 已经描述了若干示例配置,可以使用各种修改、替代构造和等同物而不脱离本公开的精神。例如,上述元素可以是更大系统的部件,其中,其他规则可以优先于或以其他方式修改本公开的各种实现或技术的应用。而且,可以在考虑上述元件之前、期间或之后进行许多步骤。

[0120] 已经提供了本申请的描述和说明,本领域技术人员可以设想到落入本申请中所讨论的总体发明概念中的变型、修改和替代实施例,其不脱离权利要求的范围。

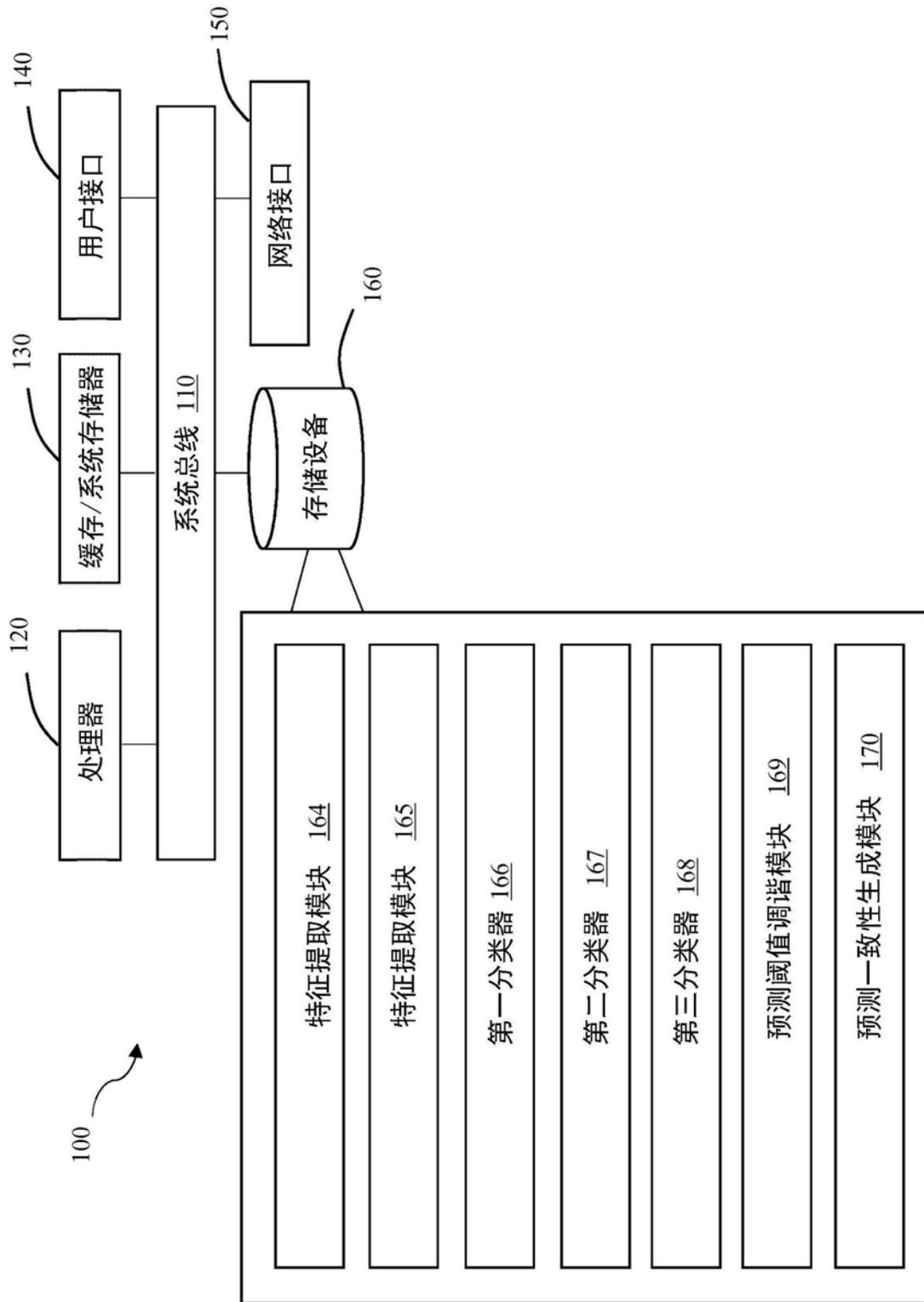


图1

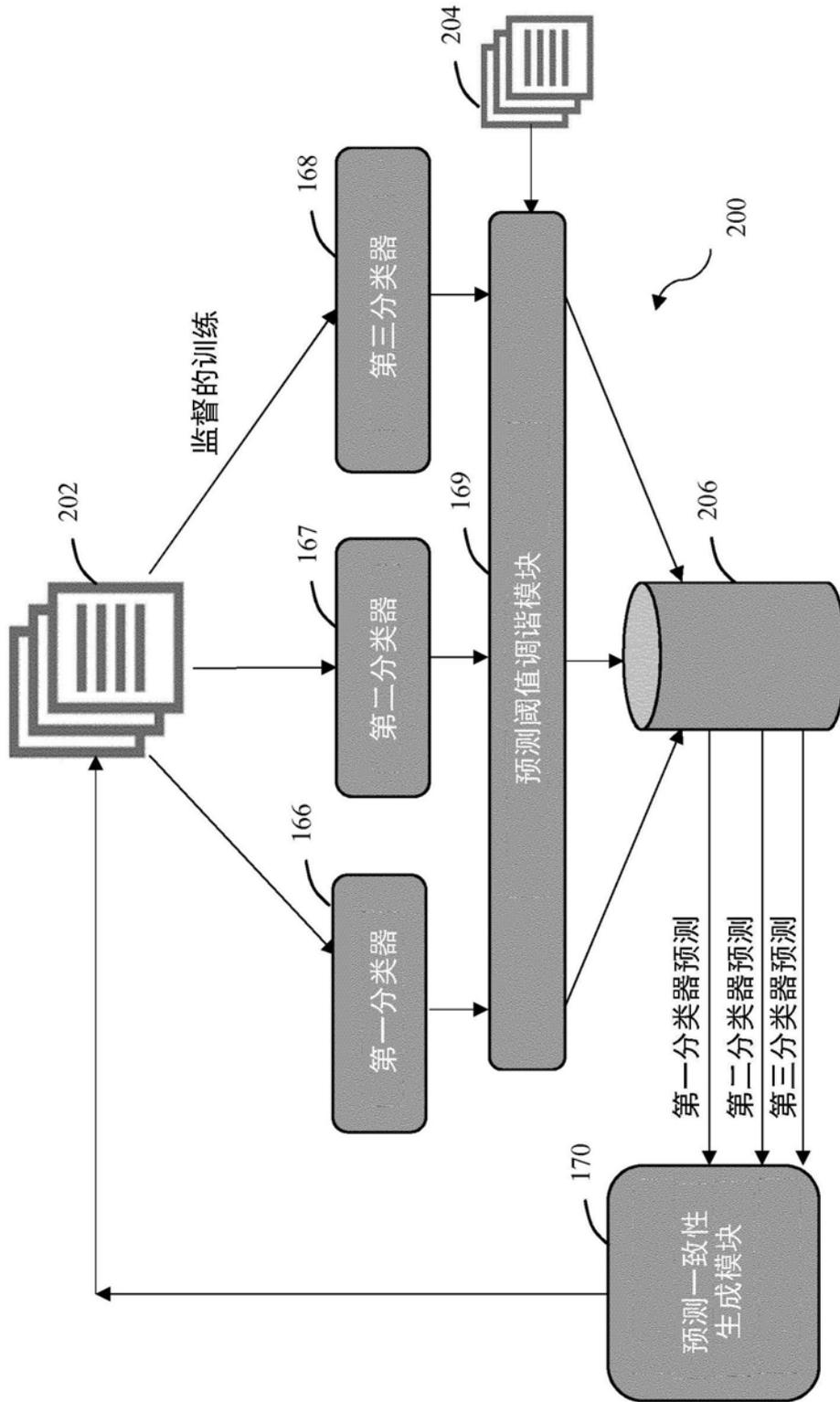


图2

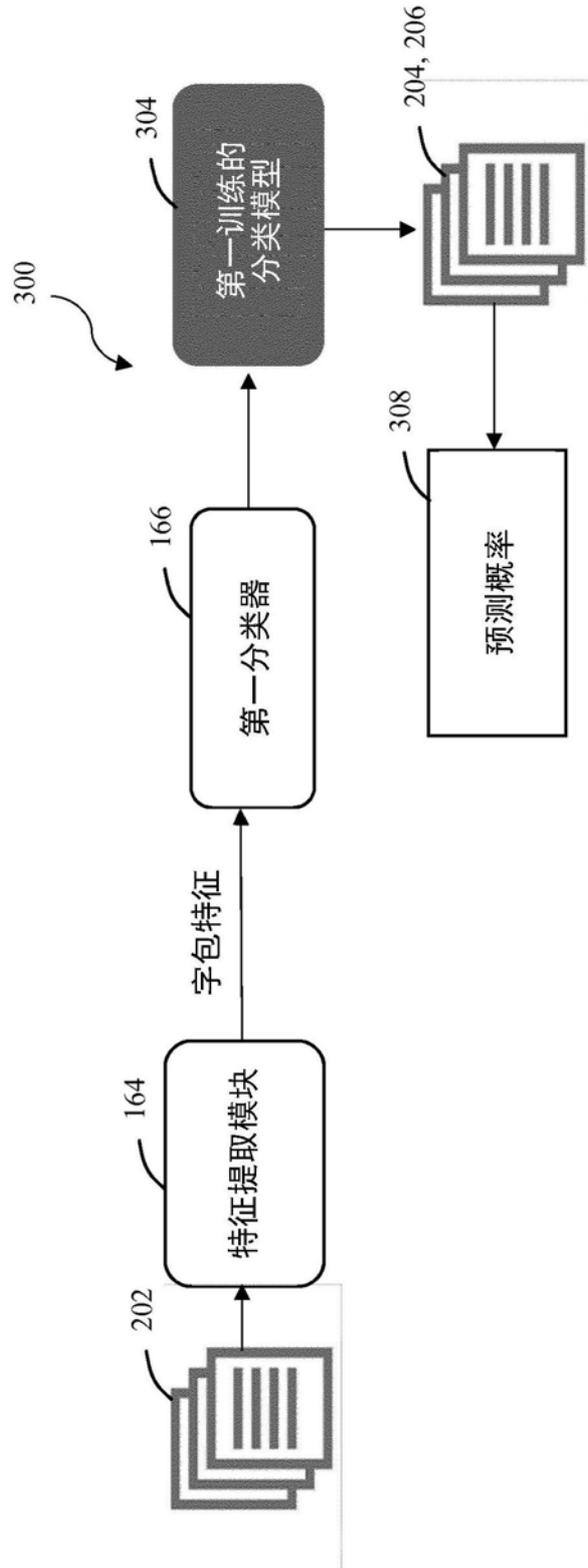


图3

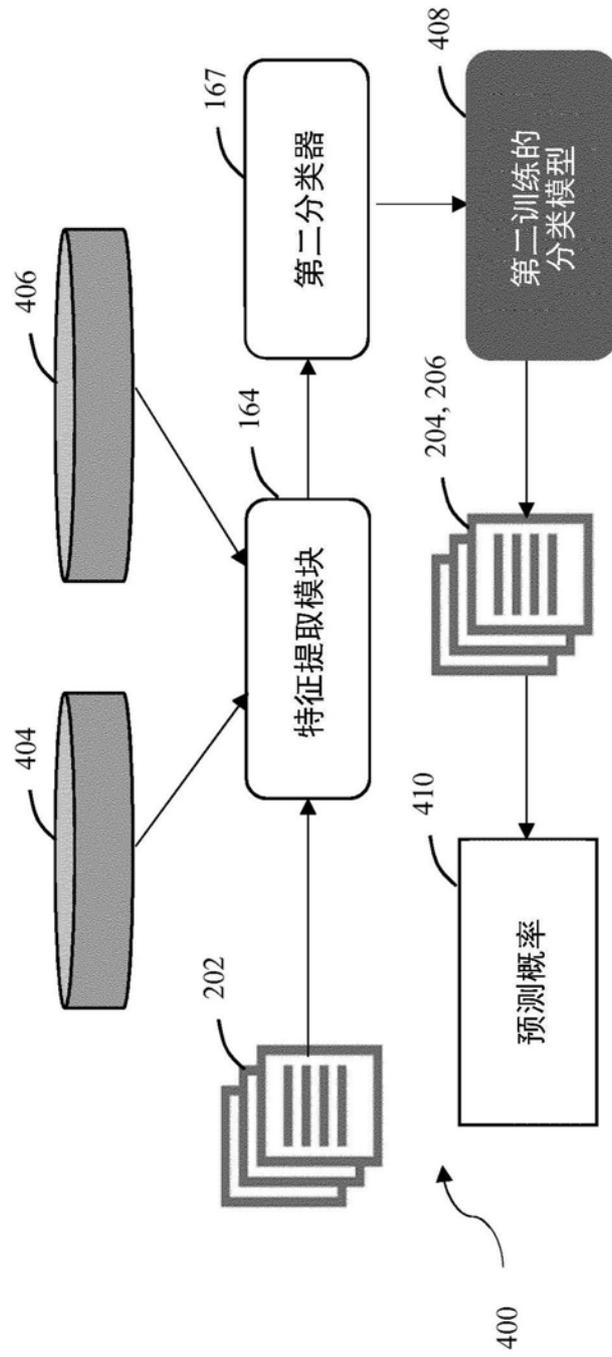


图4

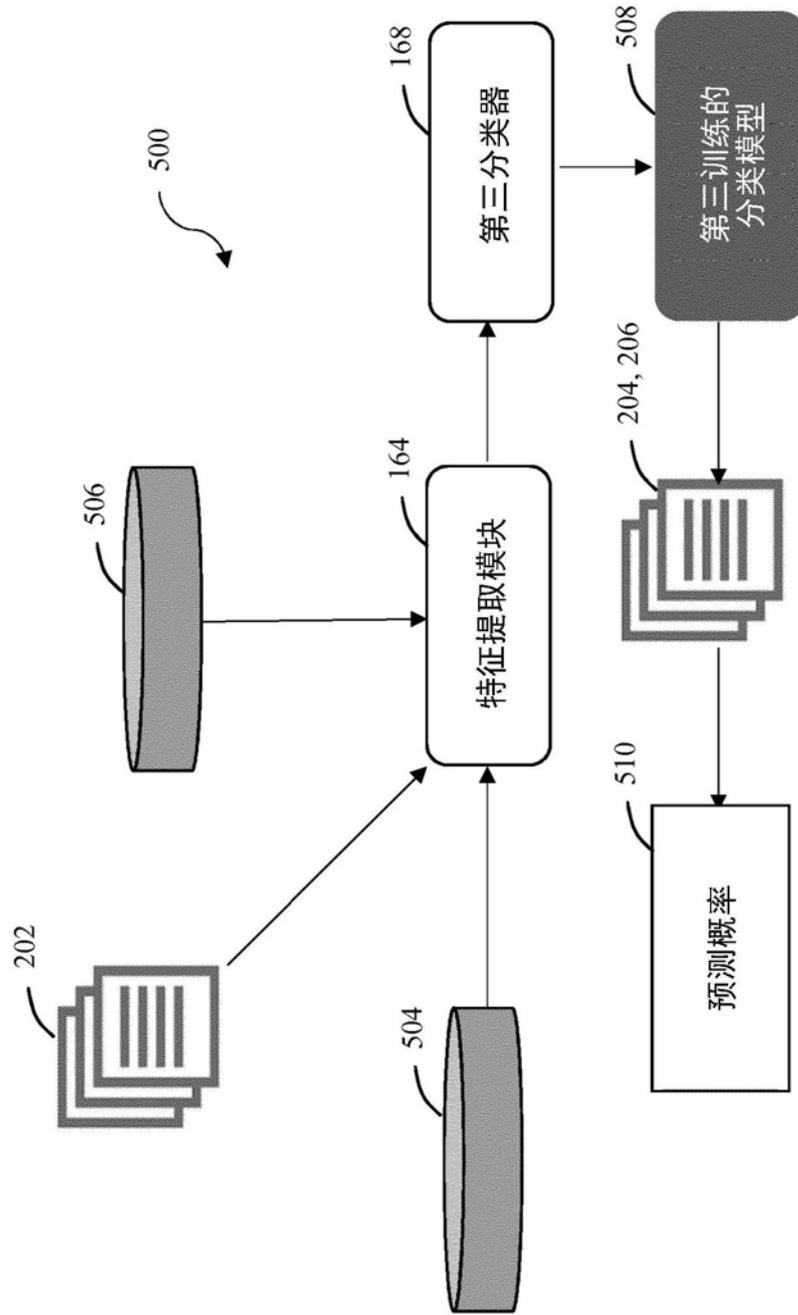


图5

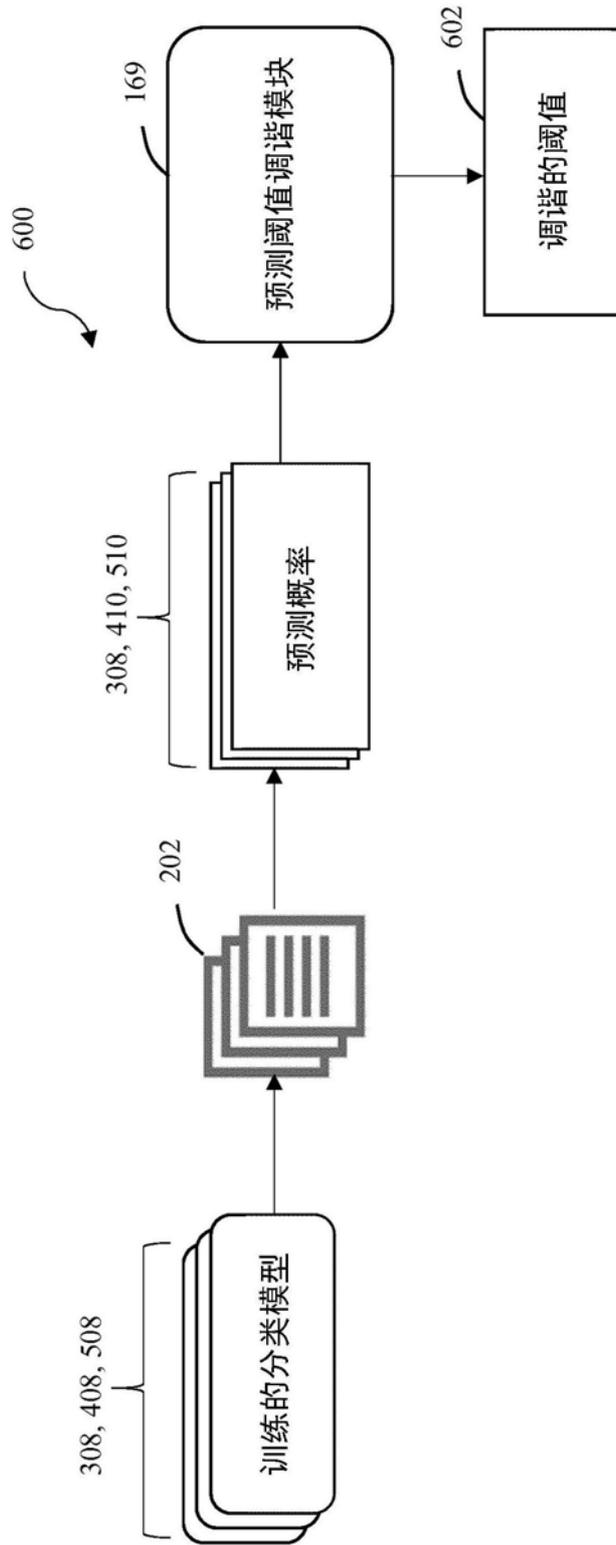


图6

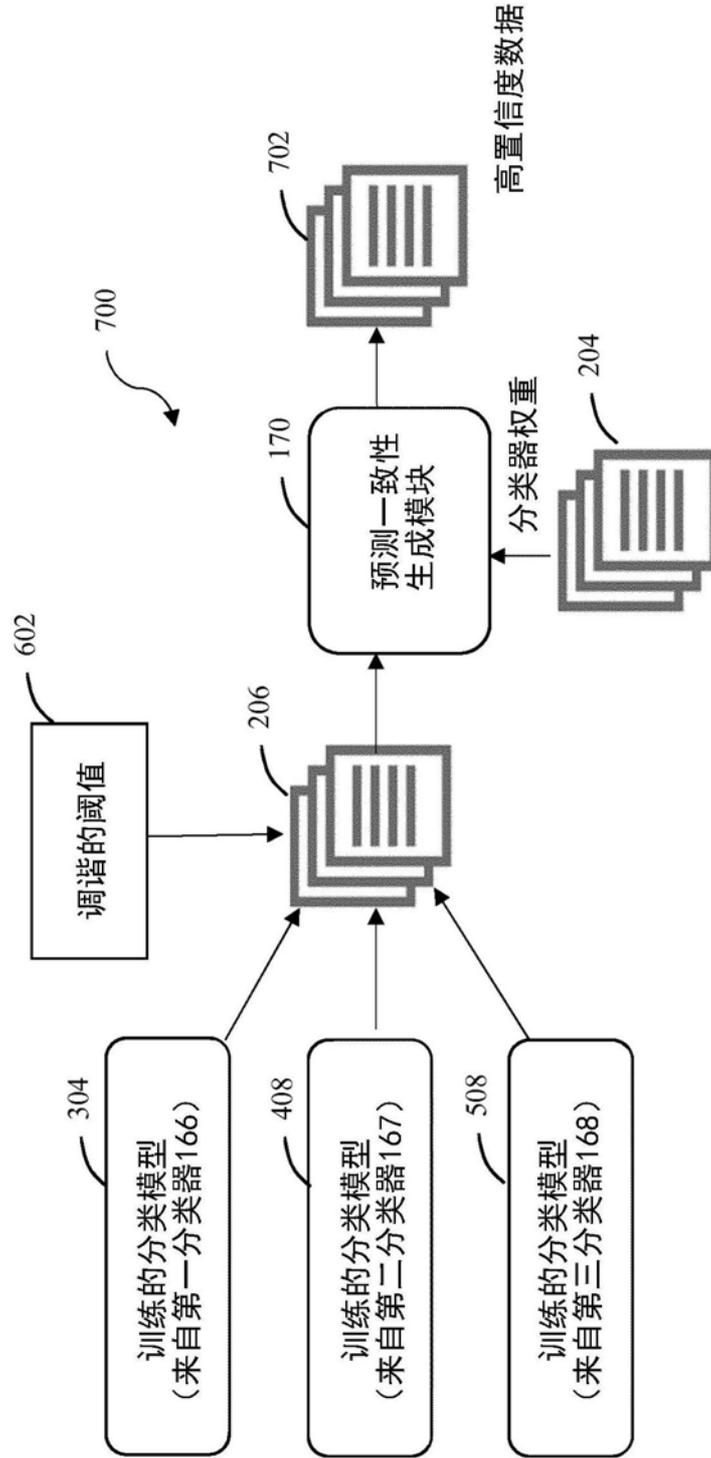


图7

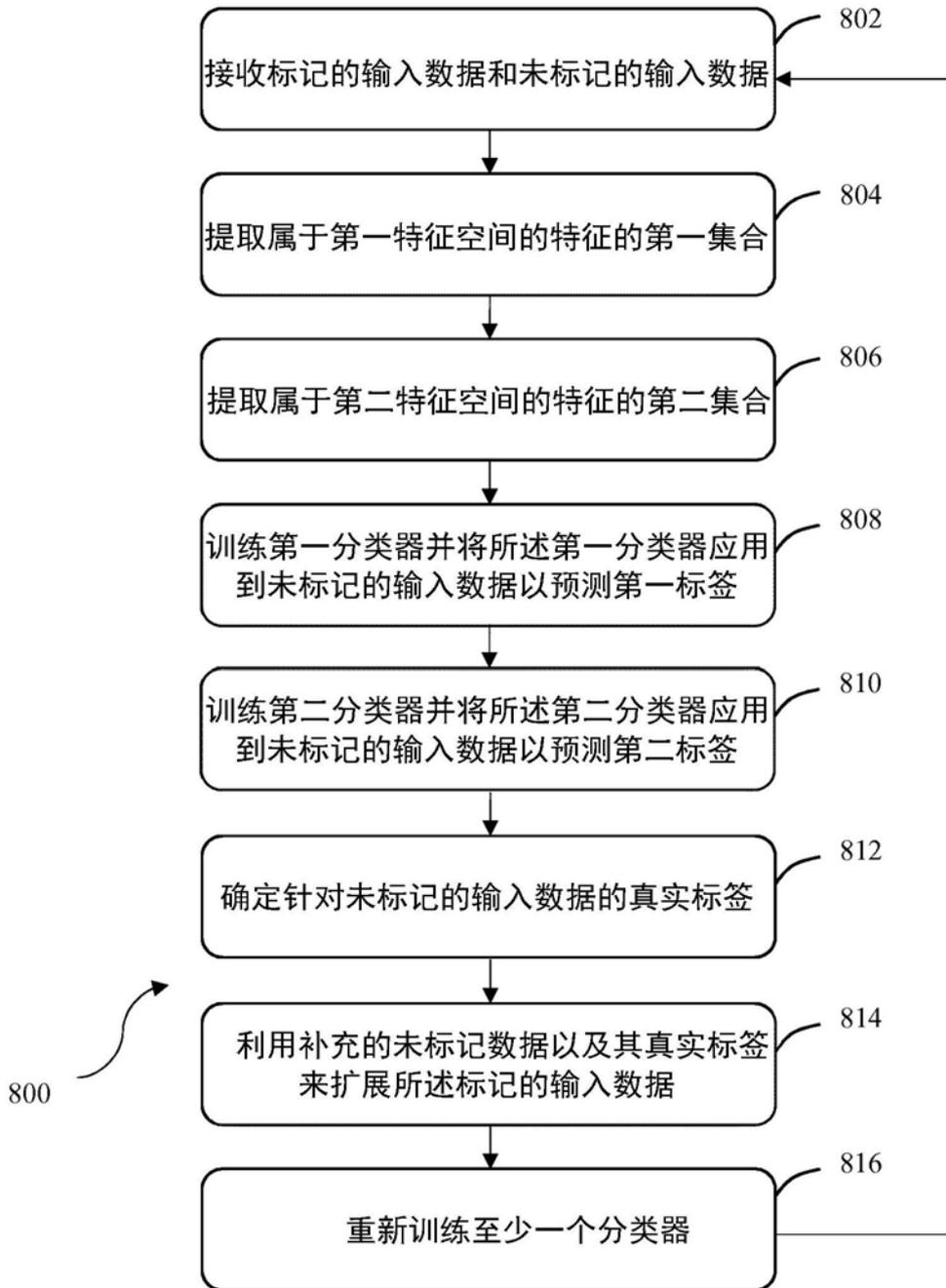


图8

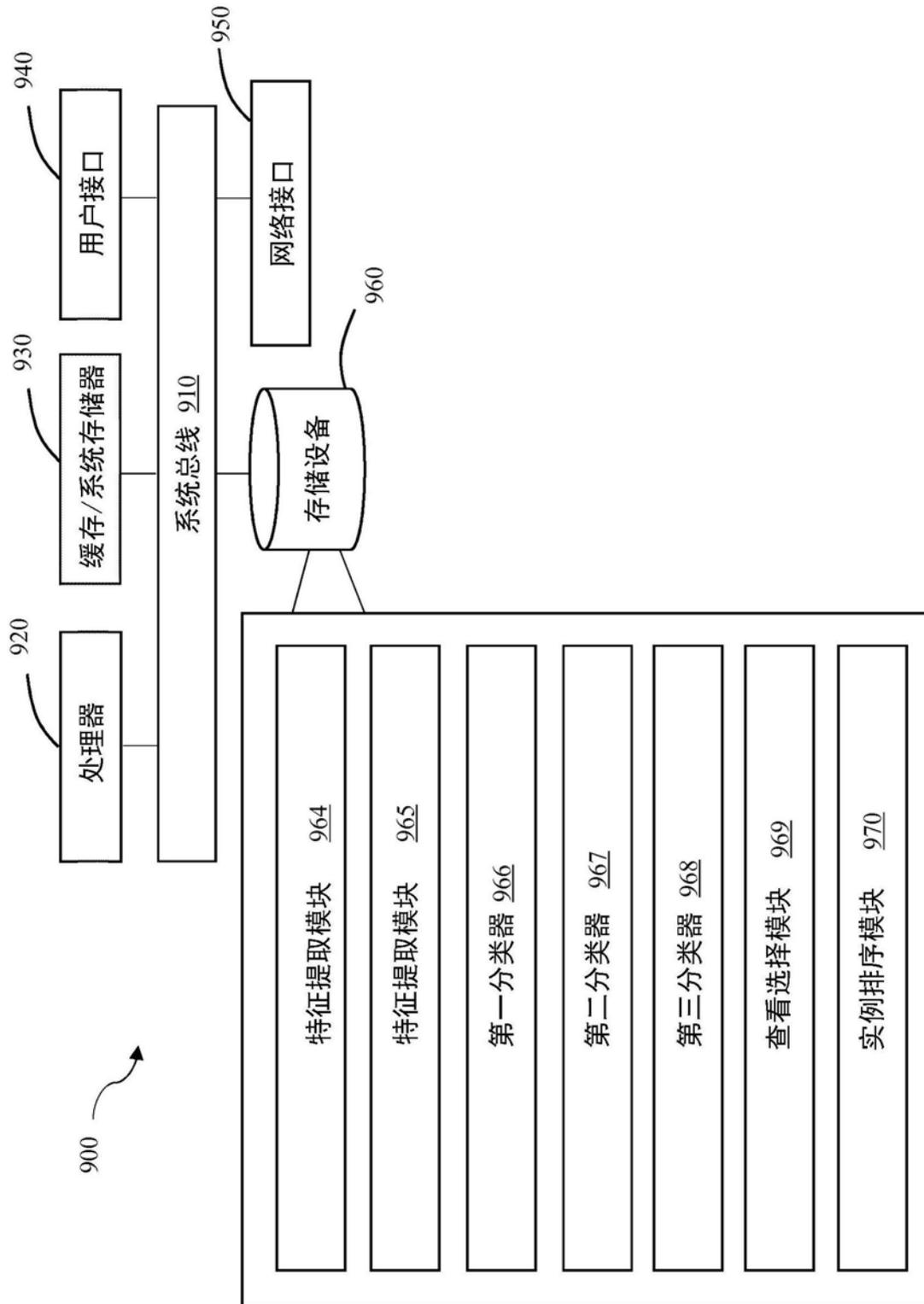


图9

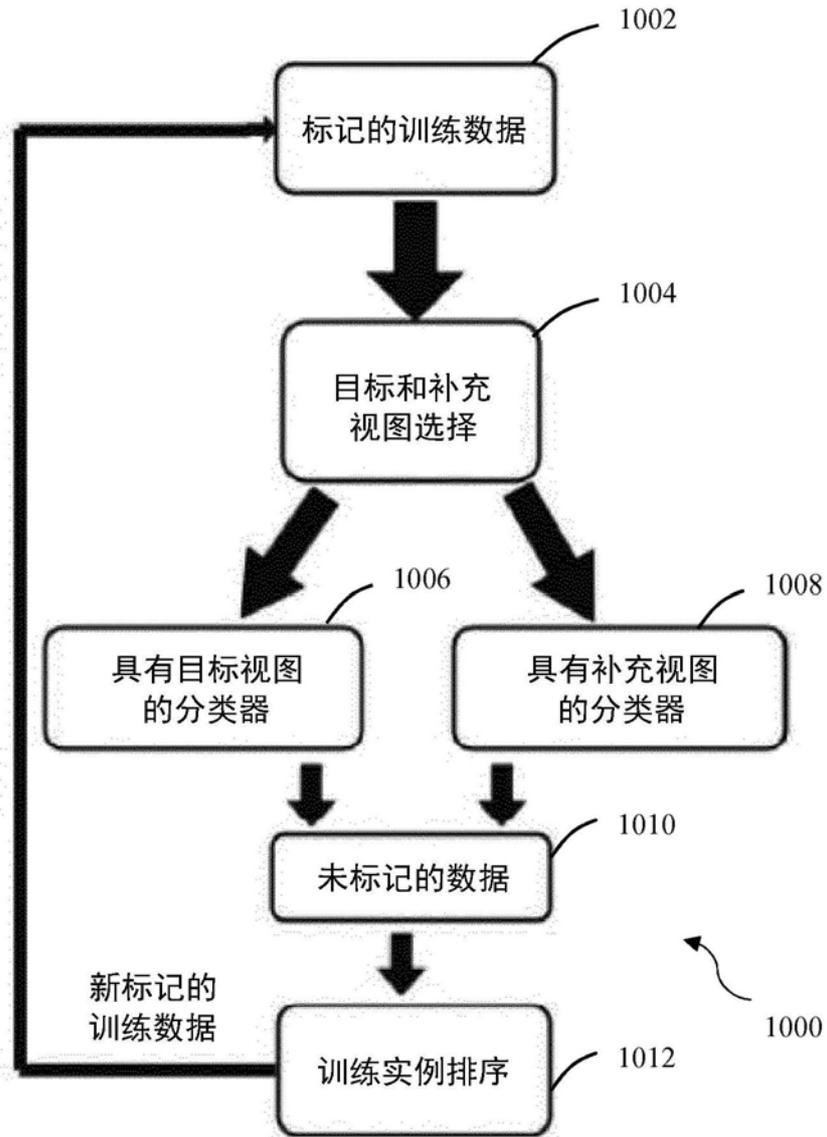


图10