(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114902239 A
(43)申请公布日 2022.08.12

(21)申请号 202080090940.8
(22)申请日 2020.08.28
(30)优先权数据
62/954,727 2019.12.30 US
(85)PCT国际申请进入国家阶段日
2022.06.29
(86)PCT国际申请的申请数据
PCT/US2020/048401 2020.08.28
(87)PCT国际申请的公布数据
WO2021/137897 EN 2021.07.08
(71)申请人 西门子公司
地址 美国新泽西州
(72)发明人 贾纳尼·韦努戈帕兰
苏迪普塔·巴萨克 夏玮
桑吉韦·斯里瓦斯塔瓦
阿伦·拉马穆尔蒂
(74)专利代理机构 北京康信知识产权代理有限责任公司 11240
专利代理师 张英
(51)Int.Cl.
G06N 3/04(2006.01)
G06N 5/00(2006.01)
G06N 5/04(2006.01)
G06N 7/00(2006.01)
G06N 5/02(2006.01)

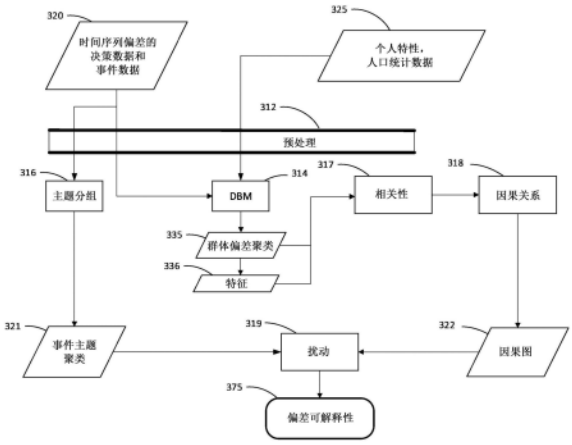
权利要求书2页 说明书11页 附图5页

(54)发明名称

深度学习模型的偏差检测和可解释性

(57)摘要

通过使用调查参与者的时间序列预测数据和事件数据以及参与者的个人特性数据，对人类决策制定进行人工智能建模来检测潜在偏差的系统和方法。深度贝叶斯模型求解偏差分布，该偏差分布将时间序列事件数据和个人特性数据的建模预测分布与由递归神经网络导出的预测概率分布拟合。针对相关个人特性的关键特征评估群体偏差聚类的集合。因果图由关键特征的依赖图定义。通过对来自因果图的特征子集的深度贝叶斯模型中的扰动来推断偏差可解释性，从而确定哪些因果关系对改变参与者的群体成员关系最敏感。



CN 114902239 A

1.一种通过人类决策制定的人工智能建模来检测潜在偏差的系统,所述系统包括:

处理器;以及

非暂时性存储器,所述非暂时性存储器上存储有由所述处理器执行的模块,所述模块包括:

时间序列事件数据的数据存储库,所述数据存储库包括由调查参与者对未来事件的预测以及事件结果,所述预测具有潜在偏差;

每个调查参与者的个人特性数据的数据存储库;

深度贝叶斯模型模块,包括:

递归神经网络,被配置用于将所述时间序列事件数据建模为预测概率分布p,

贝叶斯网络,具有至少表示估计的偏差分布的隐藏节点和表示所述个人特性数据的个人数据节点,所述贝叶斯网络被配置为接收所述概率分布并且求解将模型预测分布f最佳拟合到所述预测概率分布p的偏差分布;

聚类标识符,被配置为从所述偏差分布定义群体偏差聚类的集合;

关键特征提取器,被配置为根据所述群体偏差聚类内的共同的个人特性来识别关键特征;

相关模块,被配置为接收与所述群体偏差聚类中的每一个聚类相关的信息,并且使用依赖性分析网络来估计被识别的关键特征之间的相关性,以便基于奇异值分解为所述群体偏差聚类中的每一个聚类构建依赖图;

因果关系模块,被配置为执行因果关系分析,以使用贪婪等价搜索算法从所述依赖图中为所述群体偏差聚类中的每一个聚类导出因果图,以穿过基本图的空间来构建所述因果图,所述因果图提供了在每个群体偏差聚类和所有组合的群体偏差聚类中个人特性之间的因果关系;以及

扰动模块,被配置为通过扰动从所述因果图中导出的特征来推断偏差可解释性,以确定哪些因果关系对改变参与者的群体成员关系最敏感,其中,所述偏差可解释性包括指示哪些个人特性是基于最高灵敏度值识别的群体偏差聚类的最可能的原因。

2.根据权利要求1所述的系统,其中,所述聚类标识符函数被配置为应用曲线拟合分析,以求解将所述预测分布f最佳拟合到所述实际预测分布p的所述偏差分布,并且一旦所述曲线拟合收敛,共同检查与每个参与者相关的所述曲线拟合函数的当前参数值,以确定是否存在类似值的聚类,所述聚类用于定义群体偏差聚类的集合。

3.根据权利要求1所述的系统,其中,所述曲线拟合分析是潜在的狄利克雷分析。

4.根据权利要求1所述的系统,还包括主题模块,所述主题模块被配置为使用潜在的狄利克雷分配分析从所述时间序列事件数据中确定事件主题组;其中,所述扰动模块还被配置为包括用于推断偏差可解释性的事件主题组。

5.根据权利要求1所述的系统,其中,所述因果关系模块还被配置为执行反事实分析,以确定在所述因果图上强制执行特定边缘的效果。

6.根据权利要求1所述的系统,其中,所述因果关系模块还被配置为通过精简所述依赖图的非因果关系来导出所述因果图。

7.根据权利要求1所述的系统,其中,所述相关模块还被配置为从所述依赖图中确定多个顶部特征,所述依赖图包括表示所述特征的节点的网络,所述顶部特征是由节点相对于

其他节点的影响所定义的具有最高节点活跃度的特征，所述顶部特征被发送到所述因果关系模块用于所述因果关系分析。

8.一种通过人类决策制定的人工智能建模来检测潜在偏差的方法，所述方法包括：

通过递归神经网络将时间序列事件数据建模为预测概率分布p，其中，所述时间序列事件数据包括由调查参与者对未来事件的预测和事件结果，所述预测具有潜在偏差；

通过贝叶斯网络接收所述概率分布，所述贝叶斯网络具有至少表示估计的偏差分布的隐藏节点和表示每个调查参与者的个人特性数据的个人数据节点，并且求解将模型预测分布f最佳拟合到所述预测概率分布p的偏差分布；

从所述偏差分布定义群体偏差聚类的集合；

根据所述群体偏差聚类内的共同的个人特性来识别关键特征；

使用依赖分析网络估计被识别的所述关键特征之间的相关性，以便基于奇异值分解为所述群体偏差聚类中的每一个聚类构建依赖图；

执行因果关系分析，以使用贪婪等价搜索算法从所述依赖图中为所述群体偏差聚类中的每一个聚类导出因果图，以穿过基本图的空间来构建所述因果图，所述因果图提供了在每个群体偏差聚类和所有组合的群体偏差聚类中个人特性之间的因果关系；以及

通过扰动从所述因果图中导出的特征来推断偏差可解释性，以确定哪些因果关系对改变参与者的群体成员关系最敏感，其中，所述偏差可解释性包括指示哪些个人特性是基于最高灵敏度值识别的群体偏差聚类的最可能的原因。

9.根据权利要求8所述的方法，还包括：应用曲线拟合分析，以求解将所述预测分布f最佳拟合到所述实际预测分布p的偏差分布，并且一旦所述曲线拟合收敛，共同检查与每个参与者相关联的所述曲线拟合函数的当前参数值，以确定是否存在类似值的聚类，所述聚类用于定义群体偏差聚类的集合。

10.根据权利要求8所述的方法，其中，所述曲线拟合分析是潜在的狄利克雷分析。

11.根据权利要求8所述的方法，还包括：使用潜在的狄利克雷分配分析从所述时间序列事件数据中确定事件主题组；以及包括用于推断偏差可解释性的事件主题组。

12.根据权利要求8所述的方法，还包括：执行反事实分析，以确定在所述因果图上强制执行特定边缘的效果。

13.根据权利要求8所述的方法，还包括：通过精简所述依赖图的非因果关系来导出所述因果图。

14.根据权利要求8所述的方法，还包括：从所述依赖图中确定多个顶部特征，所述依赖图包括表示所述特征的节点的网络，所述顶部特征是由节点相对于其他节点的影响所定义的具有最高节点活跃度的特征，所述顶部特征被发送到所述因果关系模块用于所述因果关系分析。

深度学习模型的偏差检测和可解释性

技术领域

[0001] 本申请涉及深度学习模型。更具体地，本申请涉及一种系统，该系统从人类决策的深度学习模型中推断潜在群体偏差，以改进深度学习模型的可解释性。

背景技术

[0002] 近十年来，人工智能(AI)的深度学习(DL)建模分支已经彻底改变了模式识别，为在复杂动态场景中的对象提供了近乎瞬时的高质量检测和分类。当在自主系统中或作为战术决策辅助使用时，DL可以通过提高任务检测和分类的速度和质量来提高决策技能的有效性。

[0003] DL模型的可解释性有助于提高预测的置信度。例如，在训练“黑箱”DL网络后，仍不确定损失函数在寻找测试输入和已知输入之间最相似的匹配时是否表现得准确。在人类决策制定模型的领域中，不确定性的一个区域在于训练数据中存在潜在人类偏差。例如，在尝试开发用于人类预测的DL模型时，训练数据可以由数千个事件预测组成。虽然DL模型可将对预测的各种已知影响参数化以学习人类决策制定过程，但潜在方面(例如偏差)在实现完整的建模方面产生了差距。作为说明性示例，当尝试对决策进行建模时，例如针对特定任务，存在各种可能会歪曲数据驱动模型的偏差的来源。这种偏差的来源可能包括来自个人将自己视为群体一部分的特质和特性的内隐和未观察到的偏差(甚至可能是潜意识或不知不觉)，从而导致内隐偏差的群体行为。由于未观察到与共同群体特质和有偏差群体成员相关的这种内隐偏差的原因，因此尚未开发具有可解释的偏差的模型。

[0004] 在现有的工作中，传统的贝叶斯网络(BN)被用于构建表示人类认知和决策的模型。它允许专家根据生成的概率故事来指定决策过程的模型，该故事通常与人类对潜在认知过程的直觉一致。通常，BN的结构是预先指定的，并且概率模型的参数是α先验选择的。然而，在这样的模型中执行推理和学习是NP完全的(即耗时并且经常通过使用启发式方法和近似方法来解决)。这对于非常复杂的BN尤其是个问题。

[0005] 可以采用称为深度贝叶斯模型(DBM)的一类深度概率模型(DPM)来对人类决策制定进行建模。DBM的可解释性AI基于互信息、基于梯度的技术和基于相关性的分析。因此，与传统的BNs不同，没有现有的技术可以对DBMs或DBM的结果执行因果推断。

发明内容

[0006] 所公开的方法和系统可以使用因果推理和扰动分析来解决所有上述挑战，以确定决策数据中存在的潜在偏差，从而改进决策预测模型。深度贝叶斯模型(DBM)从建模的预测数据和相关联的个人特性数据中学习预测分布并识别群体偏差聚类。与群体偏差聚类相关的个人特性数据中的关键特征在完全连接的依赖图中进行关联。基于依赖图构造因果图，以识别关键特征与群体偏差聚类的因果关系。对具有因果关系的个体关键特征的扰动揭示了特征的敏感性，以便根据个人特质的特定特征对预测数据的偏差或偏好产生更稳健的相关性，从而在深度学习模型中为潜在偏差提供可解释性。

[0007] 结果模型可以通过以下方式增强可解释性：(1)提供DBM为何为个人预测某一响应的详细信息；(2)直接提供导致偏差的数据描述符(如个人特性特征)；(3)直接提供数据描述符如何关联以及哪些描述符可以产生最大响应变化的基本原理。

附图说明

[0008] 参考以下附图描述本实施例的非限制性和非穷举性的实施例，其中，除非另有说明，否则在所有附图中类似的附图标记表示类似的元件。

[0009] 图1示出了根据本发明的实施例的计算机视觉系统的示例，该系统具有对目标对象的最佳匹配3D模型的改进的检索。

[0010] 图2A示出了根据本发明的实施例的深度贝叶斯模型(DBM)的示例。

[0011] 图2B示出了图2B中所示的DBM的修改版本，用于根据本发明的实施例估计群体偏差聚类。

[0012] 图3示出了根据本发明的实施例对从决策建模中提取的可解释的潜在偏差进行建模的过程的流程图的示例。

[0013] 图4示出了可以在其中实现本发明的实施例的计算环境的示例。

具体实施方式

[0014] 所公开的方法和系统解决了在决策中理解人类偏差的问题，该问题在工业中有各种应用，例如在构建设计软件时学习偏差对用户偏好的贡献，该设计软件可以根据用户过去的数据预测或估计用户偏好。其它示例包括预测软件和根本原因推断，这会受到由过去重复的经验产生的人类偏差的影响。使用人工智能，对决策数据的观察结果进行检查，以确定模式，从而形成表明群体偏差的群体聚类，从中进行更深入的检查和扰动，可以解释哪些关键特征定义了群体，以及哪些因素会迫使成员退出群体。根据这一理解，可以提取相关性和因果关系，以检测所观察到的决策是否存在偏差或偏好。除了检测偏差外，所公开的框架还识别了哪种人类特征或特质(文化、能力、性别、教育、工作经验、技术背景等)是检测到的偏差或偏好的最大起因。举个简单的例子，在工业环境中，有利的是，了解到具有机械背景的操作人员倾向于做出可能推断故障诊断决策的决策，受技术背景偏差的影响，所做的决策会主要倾向于机械原因，这会妨碍故障排除过程。

[0015] 图1示出了根据本发明的实施例的用于数据驱动模型的系统的示例，该系统用于推断人类决策制定的改进的可解释的模型构造的群体偏差。在一个实施例中，系统包括计算设备110，该计算设备110包括处理器115和存储器111(例如，非暂时性计算机可读介质)，在存储器111上存储有各种计算机应用程序、模块或可执行程序。这些模块包括预处理模块112、局部深度贝叶斯模型(DBM)模块114、主题模块116、相关模块117、因果模块118和扰动模块119。

[0016] 本地DBM模块114是客户端模块，用于与基于云或基于web的DBM150交互，以根据有偏差的人类决策数据和事件数据120以及个人特性/人口统计数据125(即数据120的数据描述符)确定群体偏差聚类。诸如局域网(LAN)、广域网(WAN)或基于因特网的网络130与计算设备110、DBM 150和数据存储库120、125相连。

[0017] 图2A示出了根据本发明的实施例的DBM的示例。贝叶斯网络(BN)可用于建模预测，

并具有由分布定义的特性。在本发明中，应用了深度贝叶斯模型(DBM)，该模型实现深度学习(DL)网络以参数化贝叶斯网络的分布并预测参数。DBM 201包括DL网络210和BN 212，其中使用DL算法作为函数近似器来表示变量、模型参数和数据之间的一些(或全部)关系。如图2A所示，未观察到的变量p表示某个事件x的真实概率。人类专家生成“预测”的时间序列$f=f_1,...,f_{t-1}$。一系列辅助信息$n=n_0,...,n_t$(例如，新闻标题)与f一起由递归神经网络(RNN)210处理，其中$h=h_0,h_1,...,h_t$可以预测概率分布p。在DBM中，每个f_i和n_i都是随机变量，并且它们与p的关系是概率分布，由RNN 210参数化。针对该简化情况，示出了预测者的决策(预测)的贝叶斯模型212，其中年龄影响能力，而能力又反过来影响f_t。本例中的变量和参数可定义如下：

[0018] $n_t \in \mathbb{R}^N \qquad p \in (0,1) \qquad age \in (18,100)$

[0019] $x \in \{0,1\} \qquad f_t \in (0,1) \qquad competence\ (c) \in \mathbb{R}$

[0020] $P(p|n,f)=Dirichlet[\alpha=RNN(n,f)]$

[0021] $P(x=0|p)=Bernoulli(p)$

[0022] $P(c|age)=Normal[\mu=NN(age),\sigma=NN(age)]$

[0023] $P(f_t|p,c)=Normal[\mu=p,\sigma=NN(c)]$

[0024] 这些变量之间的关系是概率分布，其参数由DL神经网络211表示。通常，这种预测模型212可以另外包括多个预测者、复杂的预测者模型以及任何辅助数据(时间序列或非时间序列)。

[0025] 使用DL表示DBM的概率参数增加了模型的灵活性，并减少了它们的偏差，因为人类专家不需要将概率关系限制为简单的函数形式。通过DL算法来表达大型异构数据与可解释的模型的概率参数之间的复杂非线性关系。否则，DBM可以用作任何BN，从而可以估计某些变量的给定值、其他变量的分布以及获得最大似然值，等等。

[0026] 贝叶斯网络可通过设计来解释。用于估计函数关系的可视化DL算法非常具有挑战性。在一个实施例中，可以隐藏DL组件，同时仅向用户公开BN。可以执行DBM决策模型以产生有用的操作建议。DBM可以计算后验p(x|数据、模型)，这是给定决策模型以及任何可用数据的目标变量x的概率。根据模型和数据，x的最大d后验值是最优决策。每个建议的依据都可以通过网络追踪。例如互信息或平均因果效应(ACE)的度量量化了DBM中连接的强度。所公开的框架通过反向跟踪BN对决策节点的影响来支持其建议的可解释性。使用贝叶斯框架的主要好处之一是能够在证据方面以严格、无偏差的方式评估模型，即给出模型假设的数据的可能性。除了最简单的模型外，计算模型证据涉及解决一个困难的非分析性集成问题。传统的方法，如马尔科夫链蒙特卡洛或嵌套抽样非常耗时，通常需要针对特定任务进行调整。相比之下，具有DBM模型证据的变分推理是一类对象。在所公开的框架中，在训练期间直接优化近似模型证据。它的近似值在训练和推理过程中很容易获得。这使得所公开的框架能够支持对竞争决策模型的比较和评估。该框架使用流数据不断地重新评估多个竞争模型的证据。

[0027] 图2B示出了根据本发明的实施例的用于估计群体偏差聚类的DBM201的修改版本。DBM 220作为图2A所示DBM的变体，使用时间深度学习RNN 221对观察到的时间序列事件数据x进行建模。时间t的响应用于根据调查预测的最终结果参数化用于实际事件概率p的分布。概率分布p输入BN 222，用于偏查的行为预测。在一个实施例中，RNN 221对发生的事件

数据(例如,问题和正确选项)建模,例如调查问题$X\in x_0,\ldots,x_t$和参与者响应$Y\in y_0,\ldots,y_t$。RNN 221对给定的预测历史数据的事件的真实概率p建模。时间t的响应y_t用于参数化真实事件概率p的分布。BN222(例如,应用潜在的狄利克雷分布或分层贝叶斯模型)从RNN模型221、潜在偏差估计和个人特性数据PD中输入历史事件的概率p,以构建预测模型f_t,该模型将观察到的决策数据的分布建模为预测行为$F\in f_0,\ldots,f_t$。BN 222对估计的偏差分布建模,该分布表示随时间变化的潜在偏差,建模为隐藏节点偏差。用于偏差节点的初始分布由一个或多个先验参数θ建模。与个人特性数据PD、偏差分布偏差和事件概率p相关的分布输入预测模型f_t。这些变量之间的关系是概率分布,其参数由DL神经网络211表示。在一个实施例中,为每个类别的个人特性数据(例如,能力、性别、技术经验)建模单独的节点。估计代表偏差聚类调查参与者的特征(反映年龄、能力、教育等)的偏差分布值。在一个实施例中,应用曲线拟合分析来求解最适合预测分布f_t与实际预测分布p的偏差分布。一旦曲线拟合收敛,共同检查与每个参与者相关的曲线拟合函数的最终参数值(例如,潜在的狄利克雷分析(LDA)的参数)是否存在类似值的聚类。根据这些聚类值,定义了群体偏差聚类224。

[0028] 在一个实施例中,BN 222包括如上所述的LDA算法。传统上,LDA对于从文档中提取主题是有用的,其中文档是潜在主题的混合,每个单词都从对应于其中一个主题的分布中采样。LDA的这一功能被扩展到本发明中眼前的目标,该目标用于从对应于潜在偏差之一的分布中对每个决策进行采样。在一个实施例中,将LDA算法应用于时间序列数据320以将相关任务分组在一起,使得DBM

[0029] 图3示出了用于构建人类决策制定的可解释性模型的流程图的示例,该模型包括群体偏差推断。在一个实施例中,为特定任务或主题寻找虚拟(基于计算机的)的决策模型,其中决策对于该任务是至关重要的。对于在预测决策中具有最佳置信度的决策模型,潜在偏差或偏好将是包含的要素。所公开的框架的过程涉及基于从众多的人类预测或决策中收集的数据对给定任务域的预测或决策事件建模。根据预测/决策数据模型,可以使用深度贝叶斯模型来导出群体偏差聚类,并将其与个人特性和人口统计数据的共同关键特征相关联。进一步的处理包括因果图和灵敏度的扰动,这将为预测或决策数据中存在的潜在偏差或偏好生成可解释性模型。

[0030] 在涉及人类预测任务的实施例中,为建模收集了两种形式的数据:(1)从中发现潜在偏差的人类决策数据和事件数据,以及(2)作为用于决策数据的数据描述符收集的个人特性数据,其表征能力以及有助于发现可用于推断群体相关偏差的聚类模式的其他特质。时间序列人类决策和事件数据320可以从多个参与者的调查(例如,问题/回答格式)中收集,这些调查与未来事件的预测有关。决策和事件数据320可以捕获参与者随时间变化的预测决策,以收集对预测模型有用的未来事件相关的数据。参与者可能会被问及与用于目标任务或主题的预测或预测决策相关的问题。例如,这些问题可能涉及选项的投票,或是/否选项。每个问题可能有一个概率值(例如,“您对您的投票有多确定?”,“您的预测结果的可能性有多大?”)。一些调查可以进行很长一段时间,以产生变化分布。例如,调查可以在一年中每月重复一次,直至用于预测事件的选定日期。用于预测事件的实际结果被记录并包括在存档的时间序列事件数据320中,这对于建模预测数据和跟踪预测是否真实的概率是有用的。在一些实施例中,数据集320可以包括多达1000000到3000000个预测的数据。

[0031] 在其他的实施例中,时间序列和事件数据320与参与者在执行除预测以外的其他

类型任务时观察到的行为有关。例如,DL模型可以学习预测二进制决策以在给定情况下执行或不执行任务。在这种情况下,就影响此类决策的潜在偏差而言,寻求DL模型的可解释性。

[0032] 个人特性/人口统计数据325是用于时间序列事件数据320的数据描述符,并且可以包括一系列个人特性,例如性别、教育水平和被调查个体的能力测试得分。收集数据时的目标可能是了解文化影响(例如,食物、宗教、地区、语言),其可以识别个体的共同群体特质,通常情况下,偏差是隐含的,并且决策或预测的原因是无法观察到的。从预测数据中发现的导致隐含的偏差的其他特质的例子可以包括以下一个或多个:经验是否会改变投票行为,年龄或性别是否会影响对给定主题的预测决策,培训是否会改变以响应问题。个人特性/人口统计数据325可以表征能力并且可以用于识别偏差特质。在一方面,决策者的详细心理和认知评估(例如,大约20项措施)可能包括雷文的渐进矩阵、认知反思测试、柏林计算、希普利抽象和词汇测试分数、政治和金融知识、算术、工作记忆和类似测试分数、人口统计数据(例如性别、年龄、教育水平),自我评价(例如,责任心、经验开放性、外向性、勇气、社会价值取向、文化世界观、封闭的需要)。

[0033] 对时间序列数据320和个人特性/人口统计数据325执行数据预处理312,并且包括:(1)数据清理错误、不一致和缺失数据;(2)数据集成,以集成多个文件中的数据,并使用文件间的关系映射数据;(3)用于降维的特征提取,例如深度特征映射(例如Word2vec)和特征降维(例如PCA,tSNE);以及(4)数据转换和时态数据可视化,如规范化。

[0034] 主题分组模块316执行探索性主题数据分析,该分析生成指示用于调查问题x的事件主题组321的结果,并且可以识别用于解释任务对群体偏差聚类的影响的类似问题。与文化模型一样,假设群体偏差聚类的行为(例如,决策)将由所考虑的场景(即,与数据集上下文中的任务相关联的话题)指示。话题分组模块316使用LDA分析将相关问题和事件任务分组在一起。

[0035] DBM模块314从基于任务的模型313和个人特性/人口统计数据325接收数据,使用图2B中所述的过程,从事件数据确定预测概率p,并确定估计的群体偏差聚类335,其中事件数据x对应于时间序列数据320,并且PD对应于个人特性/人口统计数据325。在一个实施例中,DBM模块314的聚类标识符函数应用参数曲线拟合分析(例如,潜在的狄利克雷分布分析)来识别哪个参与者属于哪个群体偏差聚类,并从输入数据确定群体偏差聚类的集合。DBM模块314的关键特征提取器从群体聚类和相关数据描述符(个人特性/人口统计)中,将关键特征336识别为组中在参与者当中常见的个人特性的特征。

[0036] 由于DBM模型不是分类模型,而是受主题模型(例如潜在的狄利克雷分析(LDA))的启发,因此例如准确性、精确性召回率和曲线下面积等评估标准并不适用。对于涉及文档的主题模型,评估首先使用LDA确定每个文档的主题,然后评估所获得的主题的适当性。在DBM分析的一个实施例中,执行了类似的方法,其中具有共享个人特性特征的群体偏差聚类指示解释分组的关键特征。在一个方面,对分组执行交叉验证和“余弦相似性度量”以获得数值分数。例如,通过对90%的参与者的随机组合,将参与者分成n=50个相等的部分。使用DBM 314基于每个部分确定群体偏差模型。对于每个模型,通过个人特性数据对每个用户进行实例特征选择。使用余弦相似性为每个模型确定每个群体下的共同选定的特征。接下来,DBM 314确定由不同数据发现的相同的群体偏差聚类是否共享相似的共同特征。可通过将

群体与具有最高马修斯相关系数的群体映射来确定群体匹配。

[0037] 相关模块317获取识别的群体偏差聚类335中的每一个聚类，并通过使用依赖网络分析来估计识别的关键特征336之间的相关性，从而得到具有C_2^m连接的全连接依赖图。在一个实施例中，依赖性分析网络利用奇异值分解来计算依赖性网络的特征之间的部分相关性(例如，通过在数据集的列之间或网络节点之间执行部分相关性)。依赖性的计算基于找到依赖性网络中最高“节点活跃度”的区域，该区域由节点相对于其他网络节点的影响来定义。这些节点活跃度表示节点j对所有节点i,k∈N的成对相关性C(i,k)的平均影响。相关性影响由相关性C(i,k)和部分相关性PC之间的差值得出，如以下关系所示：

[0038] $d(i,k|j)=C(i,k)-PC(i,k|j)$

[0039] 其中i,j,k表示网络中的节点数。

[0040] 总影响D(i,j)表示节点j对节点i的总影响，定义为节点j对所有节点k的相关性C(i,k)的平均影响，表示如下：

[0041] $$D(i,j)=\frac{1}{N-1}\Sigma_{i\neq j}^{N-1}d(i,k|j)$$

[0042] 然后将节点j的节点活跃度计算为D(i,j)的和值：

[0043] $$\sum_{i\neq j}^{N-1}D(i,j)$$

[0044] 选择具有最高节点活跃度的顶部特征(例如，前10、20或50)的固定数量，并利用这些特征执行因果关系分析。

[0045] 因果关系模块318使用来自相关模块317的结果的特征子集，通过精简依赖图的非因果关系，从依赖图中为每个群体偏差聚类导出因果图322。因果图322为每个群体偏差聚类和结合的所有群体偏差聚类提供数据集中参与者特性/数据描述符(即，依赖图特征)之间的因果关系。在一个实施例中，因果关系分析使用贪婪等价搜索(GES)算法来获得因果关系并构建因果图。GES是一种基于分数的算法，它从空白图开始，分三个阶段贪婪地最大化基本(即观察)图空间中的评分函数(通常为贝叶斯信息准则(BIC)分数)：前向阶段、后向阶段和转向阶段。在前向阶段中，GES算法以与有向无环图(DAGs)空间中的单条边相加相对应的步骤在基本图的空间中移动，一旦分数不能再增加，该阶段即中止。在后向阶段，该算法执行与删除DAGs空间中的单条边相对应的移动，直到分数不能再增加为止。在转向阶段，算法执行对应于DAGs空间中单个箭头的反转的移动，直到不能再增加分数为止。GES算法在这三个阶段循环，直到不再可能增加分数。简而言之，GES算法在图形空间上最大化评分函数。由于图形空间太大，所以应用“贪婪”方法。使用GES评分进行因果关系评价的理由如下。为了估计准确的因果DAG，理论上需要保持两个关键假设：(1)因果充分性是指不存在隐藏(或潜在)的变量，并且(2)因果忠实性的定义如下：如果X_A和X_B有条件地独立于给定的X_S，则A和B在因果DAG中被S分开。然而，根据经验，如果这些假设不成立，则当节点数量不是很大时，GES算法的性能仍然是可接受的。在一个实施例中，可以在使用GES算法获得数据驱动的因果网络之前，基于专家知识预先指定因果关系。在一个实施例中，因果关系模块318被配置为另外执行反事实分析，以确定在因果图322上实施特定边缘的效果。在一个方面，可以使用图形用户界面来由用户指定边缘实施，在该图形用户界面上，可以基于使用GES算法

的观察数据来观察对网络的图形用户界面(GUI)的响应变化。

[0046] 扰动模块319细化因果关系模块318的结果,以便可以推断偏差解释性375。尽管因果图322给出了节点之间的关系,但它并没有提供关于每个节点的变化(节点敏感性)有多大足以改变调查响应和/或参与者的群体偏差聚类成员身份。为了估计变化,扰动模块319从因果图322中选择单个特征(即,确定为因果的特征子集),扰动DBM 314中所选的特征,并评估对群体成员身份的响应。如果特定特征X的扰动导致大多数群体成员的问题响应发生变化,则特征X可能成为用于该特定主题的群体偏差聚类行为的解释。偏差可解释性375指示一个或多个个人特性特征作为影响最大的因素,最有可能是决策和事件数据310中群体偏差的原因。例如,可以基于改变了群体隶属关系的群体成员的数量(例如,通过改变预测调查问题的答案),将敏感性得分分配给每个受扰动的特征。

[0047] 与文化模型一样,假设群体偏见聚类的行为将由场景指示,即与正在考虑的数据集上下文中的任务相关的主题。从因果图中导出的个体特征的扰动可以表明个体对属于给定主题的特定任务的视角或偏好的变化。在一个实施例中,该偏好改变也有助于推断偏差解释375,指示潜在偏差的另外的因素。为了检测该基于偏好的主题,扰动模块319在可解释性推断375中包括事件主题群体321。如果观察到某个特征的变化导致个人的观点或偏好在属于某个主题的大多数任务中持续变化,则该观察识别个人特性、与所考虑事件相关的主题以及与模型的偏差之间的关系,并提供用于与这些估计相关的置信度的概率估计。

[0048] 由上述建模提供的优点很多。任何应用决策或预测的领域都可以通过理解过程中潜在偏差而得到极大的改进。一旦从上述系统和过程中了解到给定任务或主题的群体偏见,任何基于计算机的决策建模都可以更好地预测结果。例如,可以改进带有应急模型的自动辅助系统的设计,例如在汽车或其他自动辅助车辆中,以预测操作员在不同操作情况和不同人口统计情况下潜在偏差或偏好。这些模型可以根据驾驶员的个人特性进行调整,例如,考虑到对这样一个人的学习偏好。其他此类决策建模应用比比皆是。

[0049] 图4示出了可以在其中实现本发明的实施例的计算环境的示例。计算环境400包括计算机系统410,该计算机系统410可以包括通信机制,例如系统总线421或用于在计算机系统410内通信信息的其他通信机制。计算机系统410还包括与系统总线421耦合的用于处理信息的一个或多个处理器420。在一个实施例中,计算环境400对应于从人类决策数据推断偏差的公开系统,其中计算机系统410涉及下文详细描述的计算机。

[0050] 处理器420可以包括一个或多个中央处理单元(CPUs)、图形处理单元(GPUs)或本领域已知的任何其他处理器。更普遍地,本文所述的处理器是用于执行存储在计算机可读介质上的机器可读指令的设备,用于执行任务,并且可以包括硬件和固件的任何一个或组合。处理器还可以包括存储可执行用于执行任务的机器可读指令的存储器。处理器通过操纵、分析、修改、转换或传输信息以供可执行过程或信息设备使用,和/或通过将信息路由到输出设备来对信息进行操作。例如,处理器可以使用或包括计算机、控制器或微处理器的功能,并且可以使用可执行指令来执行通用计算机未执行的特殊用途功能。处理器可以包括任何类型的合适的处理单元,包括但不限于中央处理单元、微处理器、精简指令集计算机(RISC)微处理器、复杂指令集计算机(CISC)微处理器、微控制器、专用集成电路(ASIC)、现场可编程逻辑门阵列(FPGA)、片上系统(SoC),数字信号处理器(DSP)等。此外,处理器420可以具有任何合适的微体系结构设计,该微体系结构设计包括任何数量的组成组件,例如寄

存器、多路复用器、算术逻辑单元、用于控制对高速缓存的读/写操作的高速缓存控制器、分支预测器等。处理器的微体系结构设计能够支持多种指令集中的任何指令集。处理器可以与任何其他的处理器耦合(电耦合和/或包括可执行组件),以实现其之间的交互和/或通信。用户界面处理器或生成器是一种已知元件,包括用于生成显示图像或其中的部分的电子电路或软件或两者的组合。用户界面包括使用户能够与处理器或其它设备交互的一个或多个显示图像。

[0051] 系统总线421可以包括系统总线、存储器总线、地址总线或消息总线中的至少一条,并且可以允许在计算机系统410的各种组件之间交换信息(例如,数据(包括计算机可执行代码)、信令等)。系统总线421可以包括但不限于存储器总线或存储器控制器、外围总线、加速图形端口等。

[0052] 继续参考图4,计算机系统410还可以包括耦合到系统总线421的系统存储器430,用于存储将由处理器420执行的信息和指令。系统存储器430可以包括易失性和/或非易失性存储器形式的计算机可读存储介质,例如只读存储器(ROM)431和/或随机存取存储器(RAM)432。RAM 432可以包括其它动态存储设备(例如,动态RAM、静态RAM和同步DRAM)。ROM 431可以包括其它的静态存储设备(例如,可编程ROM、可擦除PROM和电可擦除PROM)。此外,系统存储器430可用于在处理器420执行指令期间存储临时变量或其它的中间信息。基本输入/输出系统433(BIOS)可以存储在ROM 431中,该基本输入/输出系统433包含有助于在计算机系统410内的元件之间传输信息的基本例程,例如在启动期间。RAM 432可包含可由处理器420立即访问和/或当前正由处理器420操作的数据和/或程序模块。系统存储器430还可以包括例如操作系统434、应用模块435和其它程序模块436。应用模块435可以包括针对图1描述的前述模块,并且还可以包括用于开发应用程序的用户门户,允许输入参数进入并根据需要进行修改。

[0053] 操作系统434可以加载到存储器430中,并且可以提供在计算机系统410上执行的其他应用软件与计算机系统410的硬件资源之间的接口。更具体地,操作系统434可包括一组计算机可执行指令,用于管理计算机系统410的硬件资源并向其它的应用程序提供公共服务(例如,管理各种应用程序之间的存储器分配)。在某些示例实施例中,操作系统434可以控制被描述为存储在数据存储器440中的一个或多个程序模块的执行。操作系统434可包括现在已知或将来可被开发的任何操作系统,包括但不限于任何服务器操作系统、任何主机操作系统、或任何其它专有或非专有的操作系统。

[0054] 计算机系统410还可以包括耦合到系统总线421的磁盘/介质控制器443,以控制用于存储信息和指令的一个或多个存储设备,例如磁硬盘441和/或可移动介质驱动器442(例如,软盘驱动器、光盘驱动器、磁带驱动器、闪存驱动器和/或固态驱动器)。可以使用适当的设备接口(例如,小型计算机系统接口(SCSI)、集成设备电子设备(IDE)、通用串行总线(USB)或火线)将存储设备440添加到计算机系统410。存储设备441,442可以位于计算机系统410的外部。

[0055] 计算机系统410可以包括用于图形用户界面(GUI)461的用户输入界面460,其可以包括一个或多个输入设备,诸如键盘、触摸屏、输入板和/或定点设备,用于与计算机用户交互并向处理器420提供信息。

[0056] 计算机系统410可以响应于处理器420执行存储器(例如系统存储器430)中包含的

一个或多个指令的一个或多个序列，来执行本发明的实施例的部分或全部处理步骤。这些指令可以从存储器440的另外的计算机可读介质(例如磁硬盘441或可移动介质驱动器442)读入系统存储器430。磁硬盘441和/或可移动介质驱动器442可以包含本发明的实施例所使用的一个或多个数据存储和数据文件。数据存储440可包括但不限于数据库(例如，关系型、面向对象型等)、文件系统、平面文件，其中数据存储在计算机网络的多个节点上的分布式数据存储、对等网络数据存储等。数据存储内容和数据文件可以被加密以提高安全性。处理器420还可以用在多处理配置中，以执行包含在系统存储器430中的一个或多个指令序列。在可替换的实施例中，可使用硬接线电路代替软件指令或与软件指令组合使用。因此，实施例不限于硬件电路和软件的任何特定组合。

[0057] 如上所述，计算机系统410可以包括至少一个计算机可读介质或存储器，用于保存根据本发明的实施例编程的指令，并用于包含本文所述的数据结构、表、记录或其它数据。本文使用的术语“计算机可读介质”是指参与向处理器420提供指令以供执行的任何介质。计算机可读介质可以采用多种形式，包括但不限于非暂时性、非易失性介质、易失性介质和传输介质。非易失性介质的非限制性的示例包括光盘、固态驱动器、磁盘和磁光盘，例如磁硬盘441或可移动介质驱动器442。易失性介质的非限制性的示例包括动态存储器，例如系统存储器430。传输介质的非限制性的示例包括同轴电缆、铜线和光纤，包括构成系统总线421的导线。传输介质也可以采用声波或光波的形式，例如在无线电波和红外数据通信期间产生的声波或光波。

[0058] 用于执行本发明操作的计算机可读介质指令可以是汇编指令、指令集体系结构(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何组合编写的源代码或目标代码，包括例如Smalltalk、C++之类的面向对象的编程语言，以及例如“C”编程语言或类似编程语言的常规过程编程语言。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，部分在用户的计算机上执行，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。在后一种场景中，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)，或者连接到外部的计算机(例如，通过使用因特网服务提供商的因特网)。在一些实施例中，包括例如可编程逻辑电路、现场可编程逻辑门阵列(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个性化电子电路来执行计算机可读程序指令，以便执行本发明的各个方面。

[0059] 本文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/或框图来描述本发明的各个方面。应当理解，流程图图示和/或框图的每个框以及流程图图示和/或框图中的框的组合可以由计算机可读介质指令来实现。

[0060] 计算环境400还可以包括计算机系统410，该计算机系统使用与一个或多个远程计算机(例如远程计算设备473)的逻辑连接在网络环境中操作。网络接口470可以实现例如经由网络471与其他的远程设备473或系统和/或存储设备441、442的通信。远程计算设备473可以是个人计算机(笔记本电脑或台式机)、移动设备、服务器、路由器、网络PC、对等设备或其它的公共网络节点，并且通常包括上述相对于计算机系统410的许多或所有元件。当在联网环境中使用时，计算机系统410可以包括调制解调器472，用于在例如因特网的网络471上建立通信。调制解调器472可经由用户网络接口470或经由另外合适的机制连接到系统总线

421。

[0061] 网络471可以是本领域中通常已知的任何网络或系统，包括因特网、内联网、局域网(LAN)、广域网(WAN)、城域网(MAN)、直接连接或一系列连接、蜂窝电话网络或能够促进计算机系统410与其它计算机(例如，远程计算设备473)之间通信的任何其它网络或介质。网络471可以是有线的、无线的或其组合。有线连接可以使用以太网、通用串行总线(USB)、RJ-6或本领域中通常已知的任何其他有线连接来实现。无线连接可以使用Wi-Fi、全球微波接入互操作性(WiMAX)和蓝牙、红外、蜂窝网络、卫星或本领域中通常已知的任何其他无线连接方法来实现。此外，多个网络可以单独工作或彼此通信以促进网络471中的通信。

[0062] 应当理解，图4中描述为存储在系统存储器430中的程序模块、应用程序、计算机可执行指令、代码等仅仅是说明性的，而不是详尽的，并且描述为由任何特定模块支持的处理可以可替换地分布在多个模块中，或者由不同的模块执行。此外，可以提供各种程序模块、脚本、插件、应用程序编程接口(APIs)或本地托管在计算机系统410、远程设备473上和/或托管在可通过一个或多个网络471访问的其他计算设备上的任何其他合适的计算机可执行代码，以支持由程序模块提供的功能、应用程序或图4所示的计算机可执行的代码和/或附加或可替换的功能。此外，功能可以不同地模块化，使得被描述为由图4所示的程序模块集合共同支持的处理可以由更少或更多数量的模块执行，或者被描述为由任何特定模块支持的功能可以至少部分地由另外的模块支持。此外，支持本文所述功能的程序模块可以根据任何合适的计算模型(例如，客户端-服务器模型、对等模型等)形成可在任何数量的系统或设备上执行的一个或多个应用程序的一部分。此外，被描述为由图4所示的任何程序模块支持的任何功能可以至少部分地在硬件和/或固件中跨任意数量的设备实现。

[0063] 还应当理解，在不脱离本发明的范围的情况下，计算机系统410可以包括所描述或描绘的之外的可替换和/或附加的硬件、软件或固件组件。更具体地，应当理解，被描绘为构成计算机系统410的一部分的软件、固件或硬件组件仅仅是说明性的，并且在各种实施例中某些组件可以不存在或者可以提供附加的组件。虽然各种说明性的程序模块已被描绘和描述为存储在系统存储器430中的软件模块，但应当理解，被描述为由程序模块支持的功能可以通过硬件、软件和/或固件的任何组合来启用。应进一步了解，在各种实施例中，上述模块中的每一个可表示受支持的功能的逻辑分区。为了便于解释功能描述该逻辑分区，并且可能不代表用于实现功能的软件、硬件和/或固件的结构。因此，应当理解，在各种实施例中，描述为由特定模块提供的功能可以至少部分地由一个或多个其它模块提供。此外，在某些实施例中可能不存在一个或多个所描绘的模块，而在其它实施例中，可能存在未描绘的附加的模块，并且可支持所述功能和/或附加的功能的至少一部分。此外，虽然某些模块可被描绘和描述为另外的模块的子模块，但在某些实施例中，此类模块可被提供为独立模块或其他模块的子模块。

[0064] 虽然已经描述了本发明的具体实施例，但是本领域的普通技术人员将认识到，在本发明的范围内存在许多其它修改和可替换的实施例。例如，关于特定设备或组件所描述的任何功能和/或处理能力可以由任何其他设备或组件来执行。此外，虽然已经根据本发明实施例描述了各种说明性的实施和体系结构，但是本领域普通技术人员将理解，对本文所描述的说明性的实施和体系结构的许多其他修改也在本发明的范围内。因此，短语“基于”或其变体应解释为“至少部分基于”。

[0065] 图中的流程图和框图说明了根据本方面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体系结构、功能和操作。就此而言，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逻辑功能的一个或多个可执行指令。在一些可替换的实施中，在框中标注的功能可能出现在附图中标注的顺序之外。例如，根据所涉及的功能，连续示出的两个框实际上可以基本上同时执行，或者这些框有时可以按照相反的顺序执行。框图和/或流程图图示中的每个框以及框图和/或流程图图示中的框的组合可以由基于专用硬件的系统来实现，这些系统执行指定的功能或动作，或执行专用硬件和计算机指令的组合。

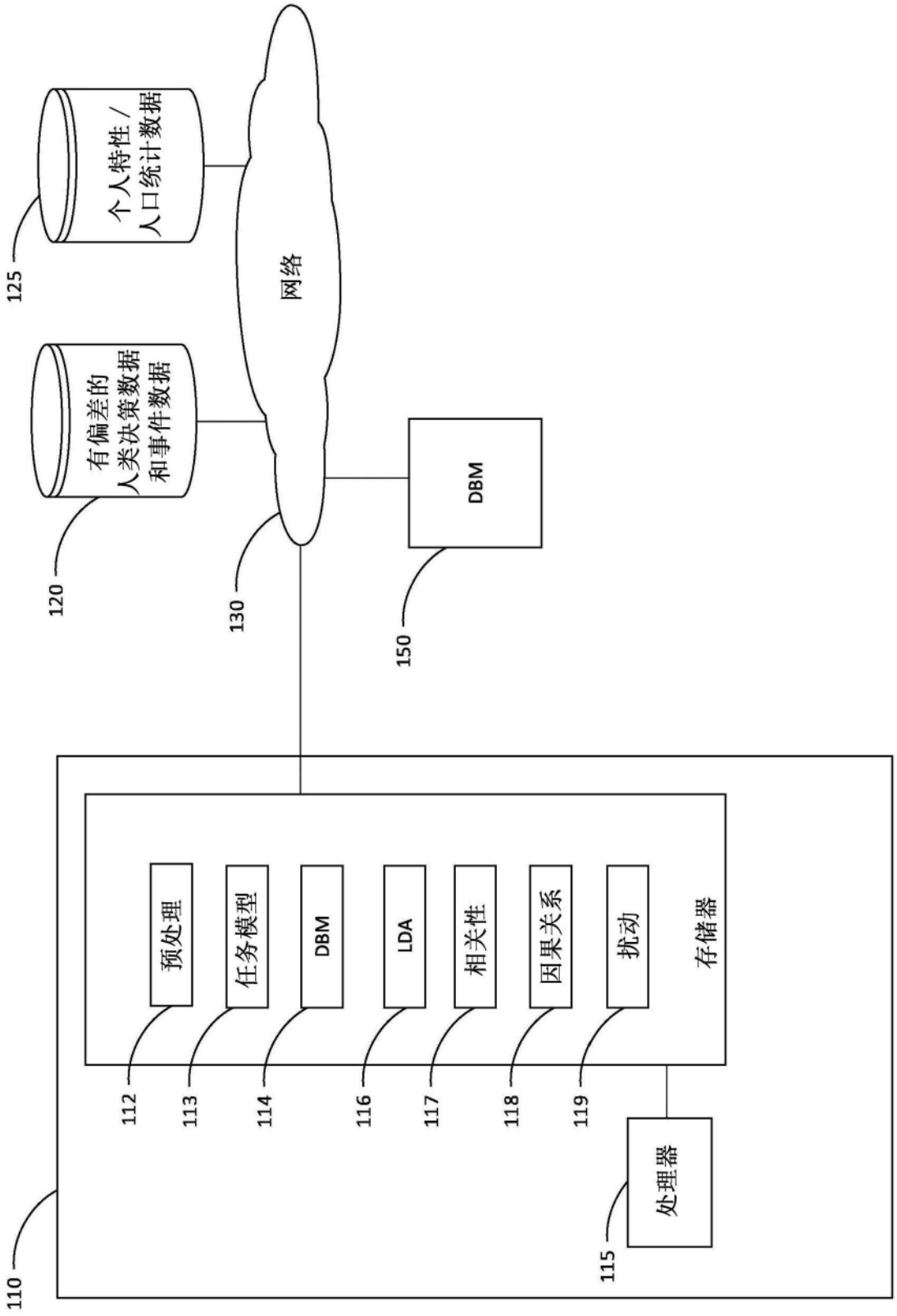

个人特性/
人口统计数据
125
网络
有偏差的
人类决策数据
和事件数据
120
130
DBM
150
预处理
任务模型
DBM
LDA
相关性
因果关系
扰动
存储器
112
113
114
116
117
118
119
处理器
115
110

图1

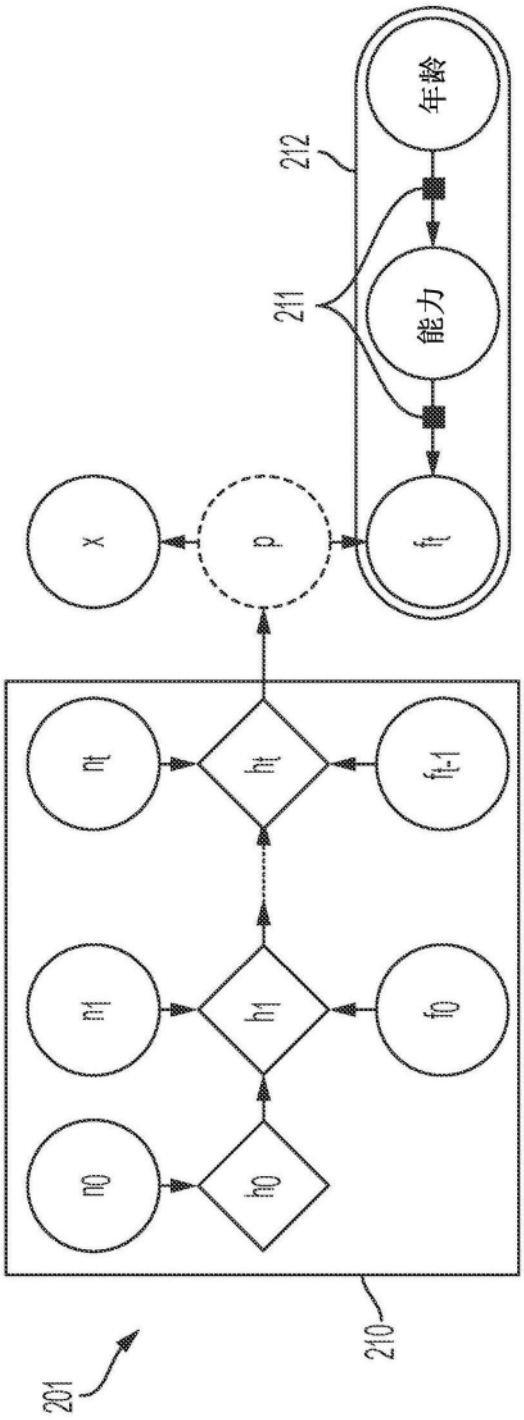


图2A

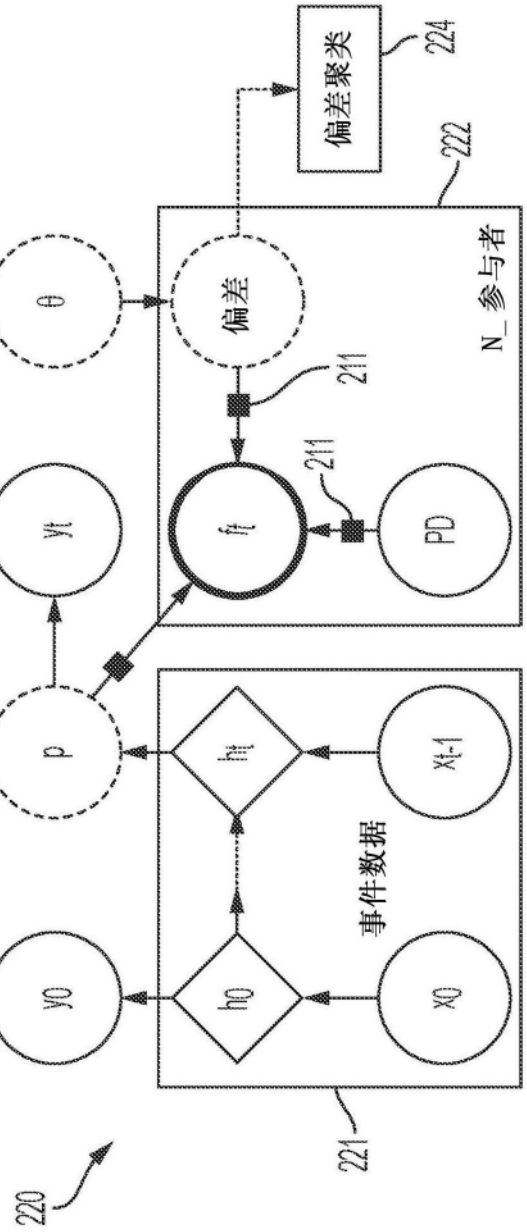

偏差聚类
224
222
θ
偏差
211
N_参与者
yt
ft
211
PD
p
hi
xt-1
事件数据
y0
h0
x0
221
220

图2B

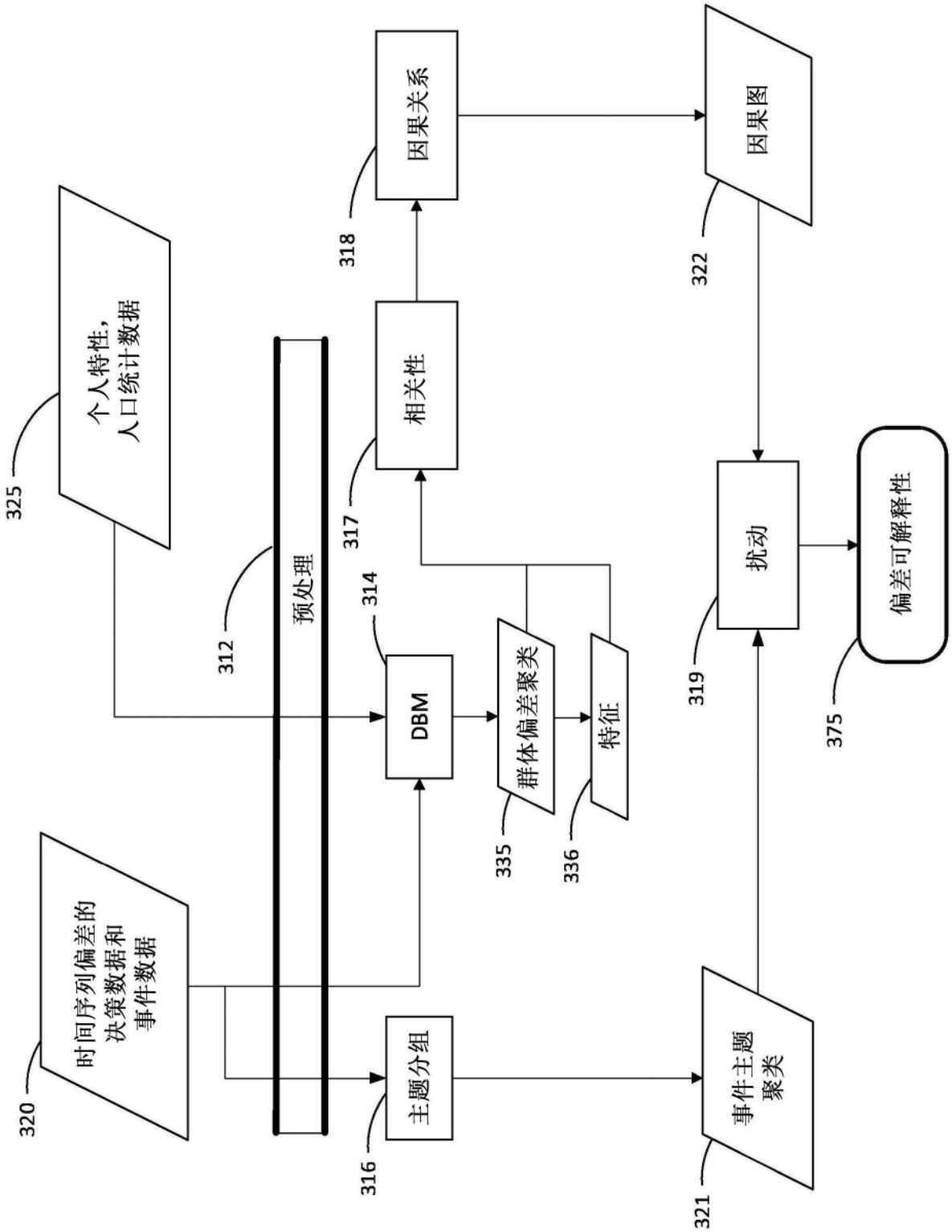

个人特性，
人口统计数据
325
预处理
312
DBM
314
相关性
317
因果关系
318
因果图
322
群体偏差聚类
335
特征
336
扰动
319
偏差可解释性
375
时间序列偏差的
决策数据和
事件数据
320
主题分组
316
事件主题
聚类
321

图3

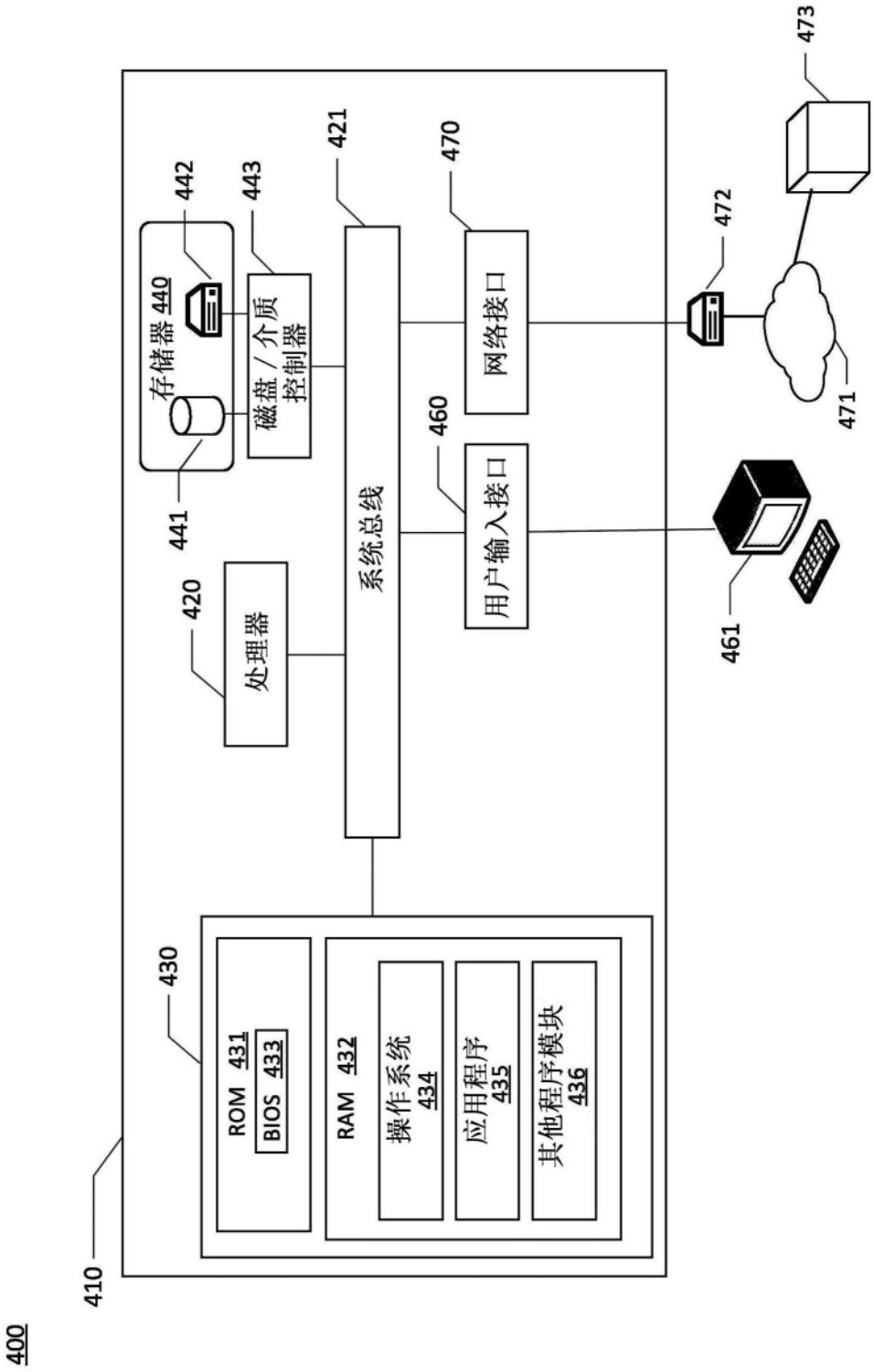

400
410
430
ROM 431
BIOS 433
RAM 432
操作系统 434
应用程序 435
其他程序模块 436
420
处理器
421
系统总线
441
存储器 440
442
443
磁盘 / 介质 控制器
460
用户输入接口
461
470
网络接口
471
472
473

图4