(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117574105 A
(43)申请公布日 2024.02.20

(21)申请号 202311513399.9

(22)申请日 2023.11.14

(71)申请人 上海富数科技有限公司
地址 200000 上海市嘉定区银翔路655号1幢4层416室

(72)发明人 尤志强 陈立峰 赵东 赵华宇 卞阳 张伟奇

(74)专利代理机构 北京慧加伦知识产权代理有限公司 16035
专利代理师 李永敏

(51)Int.Cl.
G06F 18/21(2023.01)
G06F 21/60(2013.01)

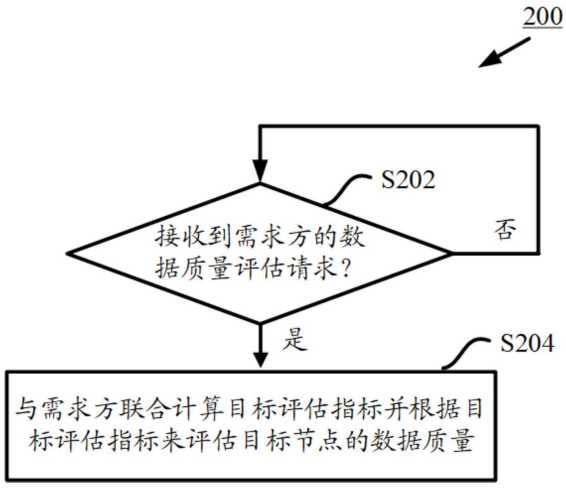
权利要求书3页 说明书15页 附图5页

(54)发明名称

对目标节点进行数据质量评估的方法及装置

(57)摘要

本公开的实施例提供一种对目标节点进行数据质量评估的方法及装置。该方法由目标节点执行。该方法包括:响应于接收到需求方的数据质量评估请求,与需求方联合计算目标评估指标并根据目标评估指标来评估目标节点的数据质量。目标评估指标包括:根据目标节点的第一数据集中的多个第一数据的方差膨胀因子来确定的第一评估指标。每个第一数据的方差膨胀因子与需求方的第二数据集相关。第二数据集包括多个第二数据。通过以下操作来计算每个第一数据的方差膨胀因子:与需求方联合计算第一数据集与第二数据集之间的皮尔逊相关系数矩阵;以及根据皮尔逊相关系数矩阵来计算第一数据的方差膨胀因子。



CN 117574105 A

1.一种对目标节点进行数据质量评估的方法,其特征在于,所述方法由所述目标节点执行,所述方法包括:

响应于接收到需求方的数据质量评估请求,与所述需求方联合计算目标评估指标并根据所述目标评估指标来评估所述目标节点的数据质量,其中,所述目标评估指标包括:根据所述目标节点的第一数据集中的多个第一数据的方差膨胀因子来确定的第一评估指标,每个第一数据的方差膨胀因子与所述需求方的第二数据集相关,所述第二数据集包括多个第二数据;

其中,通过以下操作来计算每个第一数据的方差膨胀因子:

与所述需求方联合计算所述第一数据集与所述第二数据集之间的皮尔逊相关系数矩阵,其中,所述皮尔逊相关系数矩阵包括:所述多个第一数据两两之间的第一皮尔逊相关系数,所述多个第二数据两两之间的第二皮尔逊相关系数,所述多个第一数据与所述多个第二数据两两之间的第三皮尔逊相关系数;以及

根据所述皮尔逊相关系数矩阵来计算所述第一数据的方差膨胀因子;

其中,所述第一皮尔逊相关系数由所述目标节点在本地计算,所述第二皮尔逊相关系数由所述需求方在本地计算之后发送给所述目标节点,任一第一数据与任一第二数据之间的第三皮尔逊相关系数通过以下操作来计算:

将所述第一数据中的每个特征值进行标准分数化以获得第一变换数据;

对所述第一变换数据中的每个特征值进行同态加密以获得第一加密变换数据;

向所述需求方发送所述第一加密变换数据;

接收由所述需求方生成的特征和,其中,所述特征和通过将所述第一加密变换数据中的每个加密特征值与所述第二数据中的对应特征值之积进行求和来生成;

对所接收的特征和进行同态解密以获得解密后的特征和;

响应于所述第一数据集是所述目标节点的待评估数据中的全部数据,将所述解密后的特征和除以N的商作为所述第三皮尔逊相关系数;以及

响应于所述第一数据集是所述目标节点的待评估数据中的样本数据,将所述解密后的特征和除以N-1的商作为所述第三皮尔逊相关系数;

其中,N等于所述第一数据中的特征值的个数。

2.根据权利要求1所述的方法,其特征在于,根据所述皮尔逊相关系数矩阵来计算所述第一数据的方差膨胀因子包括:

确定所述皮尔逊相关系数矩阵是否是满秩矩阵;

响应于所述皮尔逊相关系数矩阵是满秩矩阵,根据下式来计算所述第一数据的方差膨胀因子:

$$VIF_i = \frac{|M_{ii}|}{|P|}$$

其中,VIF_i表示所述第一数据集中的第i第一数据的方差膨胀因子,$|P|$表示所述皮尔逊相关系数矩阵的行列式,$|M_{ii}|$表示所述皮尔逊相关系数矩阵的第i行和第i列被去除之后的部分的行列式。

3.根据权利要求2所述的方法,其特征在于,根据所述皮尔逊相关系数矩阵来计算所述第一数据的方差膨胀因子还包括:

响应于所述皮尔逊相关系数矩阵不是满秩矩阵,根据下式来计算所述第一数据的方差膨胀因子:

$$VIF_i = \frac{1}{1 - R_i^2}$$

其中,VIF_i表示所述第一数据集中的第i第一数据的方差膨胀因子,R_i^2表示所述第i第一数据对所述第一数据集中的其他所有第一数据的多元线性回归的决定系数,其中,R_i^2根据下式来计算:

$$\begin{cases} R_i^2 = b' R_{-i,-i}^{-1} b, & Rmin < R_i^2 < Rmax \\ R_i^2 = Rmax, & R_i^2 \geq Rmax \\ R_i^2 = Rmin, & R_i^2 \leq Rmin \end{cases}$$

其中,b表示所述皮尔逊相关系数矩阵的第i列向量的第i行被去除之后的列向量,b′表示b的转置向量,$R_{-i,-i}^{-1}$表示所述皮尔逊相关系数矩阵的第i行和第i列被去除之后的子矩阵的逆矩阵,Rmax表示针对R_i^2设置的最大值,Rmin表示针对R_i^2设置的最小值。

4.根据权利要求1所述的方法,其特征在于,所述第一评估指标通过以下操作来确定:

计算所述第一数据集中方差膨胀因子超过预设阈值的第一数据的数量;以及

将所计算的数量与所述第一数据集中的所述第一数据的总数量的比值作为所述第一评估指标。

5.根据权利要求1所述的方法,其特征在于,所述目标评估指标还包括:所述第一数据集的信息价值,其中,通过以下操作来计算所述第一数据集的信息价值:

接收所述需求方发送的所述第二数据集的密态标签值和对应的唯一标识符,其中,所述密态标签值通过对所述第二数据集的标签值进行同态加密来获得;

根据所接收的唯一标识符来建立所述密态标签值与所述第一数据集的对应关系;

根据所述第一数据集对应的分箱信息以及所述密态标签值得到每一分箱的密态模糊正样本数和密态模糊负样本数;

向所述需求方发送每一分箱的密态模糊正样本数和密态模糊负样本数;

接收所述需求方发送的每一分箱的模糊证据权重、第一密态参数和第二密态参数,其中,每一分箱的所述模糊证据权重、第一密态参数和第二密态参数是根据每一分箱的密态模糊正样本数和密态模糊负样本数生成的;

根据每一分箱的所述模糊证据权重得到每一分箱的证据权重;

根据每一分箱的所述第一密态参数和所述第二密态参数得到每一分箱的密态模糊权重系数;

向所述需求方发送每一分箱的所述密态模糊权重系数;

接收所述需求方发送的每一分箱的模糊权重系数,其中,每一分箱的所述模糊权重系数是通过对每一分箱的所述密态模糊权重系数进行同态解密获得的;

根据每一分箱的所述模糊权重系数得到每一分箱的权重系数;以及

对每一分箱的所述证据权重与对应的权重系数之积进行求和以获得所述第一数据集的信息价值。

6.根据权利要求1至5中任一项所述的方法,其特征在于,所述目标评估指标还包括:对所述第一数据集的本地质量评估指标,其中,所述本地质量评估指标包括以下中的一个或多个:完整性、准确性、一致性、可用性、及时性、可理解性、异常值、可信度、关联性、数据唯一性、敏感性、数据规模、数据存储效率、数据安全性、数据历史变化程度、数据可操作性。

7.根据权利要求1至5中任一项所述的方法,其特征在于,第i第一数据与第j第一数据之间的第一皮尔逊相关系数通过以下操作来计算:

将所述第i第一数据中的每个特征值进行标准分数化以获得第i第一变换数据;

将所述第j第一数据中的每个特征值进行标准分数化以获得第j第一变换数据;

将所述第i第一变换数据中的每个特征值与所述第j第一变换数据中的对应特征值之积进行求和以获得第一和;

响应于所述第一数据集是所述目标节点的待评估数据中的全部数据,将所述第一和除以N的商作为所述第i第一数据与所述第j第一数据之间的第一皮尔逊相关系数;以及

响应于所述第一数据集是所述目标节点的待评估数据中的样本数据,将所述第一和除以N-1的商作为所述第i第一数据与所述第j第一数据之间的第一皮尔逊相关系数;

其中,N等于所述第一数据中的特征值的个数,i等于j或者i不等于j。

8.根据权利要求1至5中任一项所述的方法,其特征在于,第i第二数据与第j第二数据之间的第二皮尔逊相关系数通过以下操作来计算:

将所述第i第二数据中的每个特征值进行标准分数化以获得第i第二变换数据;

将所述第j第二数据中的每个特征值进行标准分数化以获得第j第二变换数据;

将所述第i第二变换数据中的每个特征值与所述第j第二变换数据中的对应特征值之积进行求和以获得第二和;

响应于所述第一数据集是所述目标节点的待评估数据中的全部数据,将所述第二和除以N的商作为所述第i第二数据与所述第j第二数据之间的第二皮尔逊相关系数;以及

响应于所述第一数据集是所述目标节点的待评估数据中的样本数据,将所述第二和除以N-1的商作为所述第i第二数据与所述第j第二数据之间的第二皮尔逊相关系数;

其中,N等于所述第二数据中的特征值的个数,i等于j或者i不等于j。

9.一种对目标节点进行数据质量评估的装置,其特征在于,所述装置位于所述目标节点上,所述装置包括:

至少一个处理器;以及

存储有计算机程序的至少一个存储器;

其中,当所述计算机程序由所述至少一个处理器执行时,使得所述装置执行根据权利要求1至8中任一项所述的方法的步骤。

10.一种存储有计算机程序的计算机可读存储介质,其特征在于,所述计算机程序在由处理器执行时实现根据权利要求1至8中任一项所述的方法的步骤。

对目标节点进行数据质量评估的方法及装置

技术领域

[0001] 本公开的实施例涉及计算机技术领域,具体地,涉及对目标节点进行数据质量评估的方法及装置。

背景技术

[0002] 随着互联网的发展,各类政务主体、行业主体、公司主体、机构主体可经由互联网被关联起来。每个主体可被看作一个节点。在互联网中,对数据共享的需求越来越高,数据价值的流通越来越重要。一个节点常常需要与其它节点进行数据交易或者挑选其它节点一起执行任务。在一些应用场景下,该节点需要了解其它节点相对于该节点而言的数据质量,以便进行后续操作。

发明内容

[0003] 本文中描述的实施例提供了一种对目标节点进行数据质量评估的方法、装置以及存储有计算机程序的计算机可读存储介质。目标节点可以是互联网中的一个节点。目标节点也可以是数联网中的一个节点。

[0004] 根据本公开的第一方面,提供了一种对目标节点进行数据质量评估的方法。该方法由目标节点执行。该方法包括:响应于接收到需求方的数据质量评估请求,与需求方联合计算目标评估指标并根据目标评估指标来评估目标节点的数据质量。其中,目标评估指标包括:根据目标节点的第一数据集中的多个第一数据的方差膨胀因子来确定的第一评估指标。每个第一数据的方差膨胀因子与需求方的第二数据集相关。第二数据集包括多个第二数据。其中,通过以下操作来计算每个第一数据的方差膨胀因子:与需求方联合计算第一数据集与第二数据集之间的皮尔逊相关系数矩阵,其中,皮尔逊相关系数矩阵包括:多个第一数据两两之间的第一皮尔逊相关系数,多个第二数据两两之间的第二皮尔逊相关系数,多个第一数据与多个第二数据两两之间的第三皮尔逊相关系数;以及根据皮尔逊相关系数矩阵来计算第一数据的方差膨胀因子。其中,第一皮尔逊相关系数由目标节点在本地计算。第二皮尔逊相关系数由需求方在本地计算之后发送给目标节点。任一第一数据与任一第二数据之间的第三皮尔逊相关系数通过以下操作来计算:将第一数据中的每个特征值进行标准分数化以获得第一变换数据;对第一变换数据中的每个特征值进行同态加密以获得第一加密变换数据;向需求方发送第一加密变换数据;接收由需求方生成的特征和,其中,特征和通过将第一加密变换数据中的每个加密特征值与第二数据中的对应特征值之积进行求和来生成;对所接收的特征和进行同态解密以获得解密后的特征和;响应于第一数据集是目标节点的待评估数据中的全部数据,将解密后的特征和除以N的商作为第三皮尔逊相关系数;以及响应于第一数据集是目标节点的待评估数据中的样本数据,将解密后的特征和除以N-1的商作为第三皮尔逊相关系数。其中,N等于第一数据中的特征值的个数。

[0005] 在本公开的一些实施例中,根据皮尔逊相关系数矩阵来计算第一数据的方差膨胀因子包括:确定皮尔逊相关系数矩阵是否是满秩矩阵;响应于皮尔逊相关系数矩阵是满秩

矩阵,根据下式来计算第一数据的方差膨胀因子:

[0006] $$VIF_i = \frac{|M_{ii}|}{|P|}$$

[0007] 其中,VIF_i表示第一数据集中的第i第一数据的方差膨胀因子,$|P|$表示皮尔逊相关系数矩阵的行列式,$|M_{ii}|$表示皮尔逊相关系数矩阵的第i行和第i列被去除之后的部分的行列式。

[0008] 在本公开的一些实施例中,根据皮尔逊相关系数矩阵来计算第一数据的方差膨胀因子还包括:响应于皮尔逊相关系数矩阵不是满秩矩阵,根据下式来计算第一数据的方差膨胀因子:

[0009] $$VIF_i = \frac{1}{1 - R_i^2}$$

[0010] 其中,VIF_i表示第一数据集中的第i第一数据的方差膨胀因子,R_i^2表示第i第一数据对第一数据集中的其他所有第一数据的多元线性回归的决定系数,其中,R_i^2根据下式来计算:

[0011] $$\begin{cases} R_i^2 = b' R_{-i,-i}^{-1} b, & Rmin < R_i^2 < Rmax \\ R_i^2 = Rmax, & R_i^2 \geq Rmax \\ R_i^2 = Rmin, & R_i^2 \leq Rmin \end{cases}$$

[0012] 其中,b表示皮尔逊相关系数矩阵的第i列向量的第i行被去除之后的列向量,b′表示b的转置向量,$R_{-i,-i}^{-1}$表示皮尔逊相关系数矩阵的第i行和第i列被去除之后的子矩阵的逆矩阵,Rmax表示针对R_i^2设置的最大值,Rmin表示针对R_i^2设置的最小值。

[0013] 在本公开的一些实施例中,第一评估指标通过以下操作来确定:计算第一数据集中方差膨胀因子超过预设阈值的第一数据的数量;以及将所计算的数量与第一数据集中的第一数据的总数量的比值作为第一评估指标。

[0014] 在本公开的一些实施例中,目标评估指标还包括:第一数据集的信息价值。其中,通过以下操作来计算第一数据集的信息价值:接收需求方发送的第二数据集的密态标签值和对应的唯一标识符,其中,密态标签值通过对第二数据集的标签值进行同态加密来获得;根据所接收的唯一标识符来建立密态标签值与第一数据集的对应关系;根据第一数据集对应的分箱信息以及密态标签值得到每一分箱的密态模糊正样本数和密态模糊负样本数;向需求方发送每一分箱的密态模糊正样本数和密态模糊负样本数;接收需求方发送的每一分箱的模糊证据权重、第一密态参数和第二密态参数,其中,每一分箱的模糊证据权重、第一密态参数和第二密态参数是根据每一分箱的密态模糊正样本数和密态模糊负样本数生成的;根据每一分箱的模糊证据权重得到每一分箱的证据权重;根据每一分箱的第一密态参数和第二密态参数得到每一分箱的密态模糊权重系数;向需求方发送每一分箱的密态模糊权重系数;接收需求方发送的每一分箱的模糊权重系数,其中,每一分箱的模糊权重系数是通过对每一分箱的密态模糊权重系数进行同态解密获得的;根据每一分箱的模糊权重系数得到每一分箱的权重系数;以及对每一分箱的证据权重与对应的权重系数之积进行求和以获得第一数据集的信息价值。

[0015] 在本公开的一些实施例中,目标评估指标还包括:对第一数据集的本地质量评估指标。其中,本地质量评估指标包括以下中的一个或多个:完整性、准确性、一致性、可用性、及时性、可理解性、异常值、可信度、关联性、数据唯一性、敏感性、数据规模、数据存储效率、数据安全性、数据历史变化程度、数据可操作性。

[0016] 在本公开的一些实施例中,第i第一数据与第j第一数据之间的第一皮尔逊相关系数通过以下操作来计算:将第i第一数据中的每个特征值进行标准分数化以获得第i第一变换数据;将第j第一数据中的每个特征值进行标准分数化以获得第j第一变换数据;将第i第一变换数据中的每个特征值与第j第一变换数据中的对应特征值之积进行求和以获得第一和;响应于第一数据集是目标节点的待评估数据中的全部数据,将第一和除以N的商作为第i第一数据与第j第一数据之间的第一皮尔逊相关系数;以及响应于第一数据集是目标节点的待评估数据中的样本数据,将第一和除以N-1的商作为第i第一数据与第j第一数据之间的第一皮尔逊相关系数。其中,N等于第一数据中的特征值的个数。i等于j或者i不等于j。

[0017] 在本公开的一些实施例中,第i第二数据与第j第二数据之间的第二皮尔逊相关系数通过以下操作来计算:将第i第二数据中的每个特征值进行标准分数化以获得第i第二变换数据;将第j第二数据中的每个特征值进行标准分数化以获得第j第二变换数据;将第i第二变换数据中的每个特征值与第j第二变换数据中的对应特征值之积进行求和以获得第二和;响应于第一数据集是目标节点的待评估数据中的全部数据,将第二和除以N的商作为第i第二数据与第j第二数据之间的第二皮尔逊相关系数;以及响应于第一数据集是目标节点的待评估数据中的样本数据,将第二和除以N-1的商作为第i第二数据与第j第二数据之间的第二皮尔逊相关系数。其中,N等于第二数据中的特征值的个数。i等于j或者i不等于j。

[0018] 根据本公开的第二方面,提供了一种对目标节点进行数据质量评估的装置。该装置位于目标节点上。该装置包括至少一个处理器;以及存储有计算机程序的至少一个存储器。当计算机程序由至少一个处理器执行时,使得装置响应于接收到需求方的数据质量评估请求,与需求方联合计算目标评估指标并根据目标评估指标来评估目标节点的数据质量。其中,目标评估指标包括:根据目标节点的第一数据集中的多个第一数据的方差膨胀因子来确定的第一评估指标。每个第一数据的方差膨胀因子与需求方的第二数据集相关。第二数据集包括多个第二数据。其中,计算机程序在由至少一个处理器执行时使得装置通过以下操作来计算每个第一数据的方差膨胀因子:与需求方联合计算第一数据集与第二数据集之间的皮尔逊相关系数矩阵,其中,皮尔逊相关系数矩阵包括:多个第一数据两两之间的第一皮尔逊相关系数,多个第二数据两两之间的第二皮尔逊相关系数,多个第一数据与多个第二数据两两之间的第三皮尔逊相关系数;以及根据皮尔逊相关系数矩阵来计算第一数据的方差膨胀因子。其中,第一皮尔逊相关系数由目标节点在本地计算。第二皮尔逊相关系数由需求方在本地计算之后发送给目标节点。任一第一数据与任一第二数据之间的第三皮尔逊相关系数通过以下操作来计算:将第一数据中的每个特征值进行标准分数化以获得第一变换数据;对第一变换数据中的每个特征值进行同态加密以获得第一加密变换数据;向需求方发送第一加密变换数据;接收由需求方生成的特征和,其中,特征和通过将第一加密变换数据中的每个加密特征值与第二数据中的对应特征值之积进行求和来生成;对所接收的特征和进行同态解密以获得解密后的特征和;响应于第一数据集是目标节点的待评估数据中的全部数据,将解密后的特征和除以N的商作为第三皮尔逊相关系数;以及响应于第一

数据集是目标节点的待评估数据中的样本数据,将解密后的特征和除以N-1的商作为第三皮尔逊相关系数。其中,N等于第一数据中的特征值的个数。

[0019] 在本公开的一些实施例中,计算机程序在由至少一个处理器执行时使得装置通过以下操作来根据皮尔逊相关系数矩阵来计算第一数据的方差膨胀因子:确定皮尔逊相关系数矩阵是否是满秩矩阵;响应于皮尔逊相关系数矩阵是满秩矩阵,根据下式来计算第一数据的方差膨胀因子:

[0020] $$VIF_i = \frac{|M_{ii}|}{|P|}$$

[0021] 其中,VIF_i表示第一数据集中的第i第一数据的方差膨胀因子,$|P|$表示皮尔逊相关系数矩阵的行列式,$|M_{ii}|$表示皮尔逊相关系数矩阵的第i行和第i列被去除之后的部分的行列式。

[0022] 在本公开的一些实施例中,计算机程序在由至少一个处理器执行时使得装置还通过以下操作来根据皮尔逊相关系数矩阵来计算第一数据的方差膨胀因子:响应于皮尔逊相关系数矩阵不是满秩矩阵,根据下式来计算第一数据的方差膨胀因子:

[0023] $$VIF_i = \frac{1}{1 - R_i^2}$$

[0024] 其中,VIF_i表示第一数据集中的第i第一数据的方差膨胀因子,R_i^2表示第i第一数据对第一数据集中的其他所有第一数据的多元线性回归的决定系数,其中,R_i^2根据下式来计算:

[0025] $$\begin{cases} R_i^2 = b' R_{-i,-i}^{-1} b, & Rmin < R_i^2 < Rmax \\ R_i^2 = Rmax, & R_i^2 \geq Rmax \\ R_i^2 = Rmin, & R_i^2 \leq Rmin \end{cases}$$

[0026] 其中,b表示皮尔逊相关系数矩阵的第i列向量的第i行被去除之后的列向量,b′表示b的转置向量,$R_{-i,-i}^{-1}$表示皮尔逊相关系数矩阵的第i行和第i列被去除之后的子矩阵的逆矩阵,Rmax表示针对R_i^2设置的最大值,Rmin表示针对R_i^2设置的最小值。

[0027] 在本公开的一些实施例中,计算机程序在由至少一个处理器执行时使得装置通过以下操作来确定第一评估指标:计算第一数据集中方差膨胀因子超过预设阈值的第一数据的数量;以及将所计算的数量与第一数据集中的第一数据的总数量的比值作为第一评估指标。

[0028] 在本公开的一些实施例中,目标评估指标还包括:第一数据集的信息价值。计算机程序在由至少一个处理器执行时使得装置通过以下操作来计算第一数据集的信息价值:接收需求方发送的第二数据集的密态标签值和对应的唯一标识符,其中,密态标签值通过对第二数据集的标签值进行同态加密来获得;根据所接收的唯一标识符来建立密态标签值与第一数据集的对应关系;根据第一数据集对应的分箱信息以及密态标签值得到每一分箱的密态模糊正样本数和密态模糊负样本数;向需求方发送每一分箱的密态模糊正样本数和密态模糊负样本数;接收需求方发送的每一分箱的模糊证据权重、第一密态参数和第二密态参数,其中,每一分箱的模糊证据权重、第一密态参数和第二密态参数是根据每一分箱的密

态模糊正样本数和密态模糊负样本数生成的;根据每一分箱的模糊证据权重得到每一分箱的证据权重;根据每一分箱的第一密态参数和第二密态参数得到每一分箱的密态模糊权重系数;向需求方发送每一分箱的密态模糊权重系数;接收需求方发送的每一分箱的模糊权重系数,其中,每一分箱的模糊权重系数是通过对每一分箱的密态模糊权重系数进行同态解密获得的;根据每一分箱的模糊权重系数得到每一分箱的权重系数;以及对每一分箱的证据权重与对应的权重系数之积进行求和以获得第一数据集的信息价值。

[0029] 在本公开的一些实施例中,计算机程序在由至少一个处理器执行时使得装置通过以下操作来计算第i第一数据与第j第一数据之间的第一皮尔逊相关系数:将第i第一数据中的每个特征值进行标准分数化以获得第i第一变换数据;将第j第一数据中的每个特征值进行标准分数化以获得第j第一变换数据;将第i第一变换数据中的每个特征值与第j第一变换数据中的对应特征值之积进行求和以获得第一和;响应于第一数据集是目标节点的待评估数据中的全部数据,将第一和除以N的商作为第i第一数据与第j第一数据之间的第一皮尔逊相关系数;以及响应于第一数据集是目标节点的待评估数据中的样本数据,将第一和除以N-1的商作为第i第一数据与第j第一数据之间的第一皮尔逊相关系数。其中,N等于第一数据中的特征值的个数。i等于j或者i不等于j。

[0030] 在本公开的一些实施例中,计算机程序在由至少一个处理器执行时使得装置通过以下操作来计算第i第二数据与第j第二数据之间的第二皮尔逊相关系数:将第i第二数据中的每个特征值进行标准分数化以获得第i第二变换数据;将第j第二数据中的每个特征值进行标准分数化以获得第j第二变换数据;将第i第二变换数据中的每个特征值与第j第二变换数据中的对应特征值之积进行求和以获得第二和;响应于第一数据集是目标节点的待评估数据中的全部数据,将第二和除以N的商作为第i第二数据与第j第二数据之间的第二皮尔逊相关系数;以及响应于第一数据集是目标节点的待评估数据中的样本数据,将第二和除以N-1的商作为第i第二数据与第j第二数据之间的第二皮尔逊相关系数。其中,N等于第二数据中的特征值的个数。i等于j或者i不等于j。

[0031] 根据本公开的第三方面,提供了一种存储有计算机程序的计算机可读存储介质,其中,计算机程序在由处理器执行时实现根据本公开的第一方面所述的方法的步骤。

附图说明

[0032] 为了更清楚地说明本公开的实施例的技术方案,下面将对实施例的附图进行简要说明,应当知道,以下描述的附图仅仅涉及本公开的一些实施例,而非对本公开的限制,其中:

[0033] 图1是数联网的示意性拓扑图;

[0034] 图2是根据本公开的实施例的对目标节点进行数据质量评估的方法的示意性流程图;

[0035] 图3是根据本公开的实施例的计算第一数据的方差膨胀因子的步骤的示意性流程图;

[0036] 图4是皮尔逊相关系数矩阵的示例图;

[0037] 图5是根据本公开的实施例的计算任一第一数据与任一第二数据之间的第三皮尔逊相关系数的步骤的示意性流程图;

[0038] 图6是根据本公开的实施例的对目标节点进行数据质量评估的装置的示意性框图。

[0039] 在附图中,最后两位数字相同的标记对应于相同的元素。需要注意的是,附图中的元素是示意性的,没有按比例绘制。

具体实施方式

[0040] 为了使本公开的实施例的目的、技术方案和优点更加清楚,下面将结合附图,对本公开的实施例的技术方案进行清楚、完整的描述。显然,所描述的实施例是本公开的一部分实施例,而不是全部的实施例。基于所描述的本公开的实施例,本领域技术人员在无需创造性劳动的前提下所获得的所有其它实施例,也都属于本公开保护的范围。

[0041] 除非另外定义,否则在此使用的所有术语(包括技术和科学术语)具有与本公开主题所属领域的技术人员所通常理解的相同含义。进一步将理解的是,诸如在通常使用的词典中定义的那些的术语应解释为具有与说明书上下文和相关技术中它们的含义一致的含义,并且将不以理想化或过于正式的形式来解释,除非在此另外明确定义。另外,诸如“第一”和“第二”的术语仅用于将一个部件(或部件的一部分)与另一个部件(或部件的另一部分)区分开。

[0042] 图1示出数联网的示意性拓扑图。数联网可包括多个子网10。每个子网10包括枢纽节点11和与枢纽节点直接连接的多个参与节点12。该多个子网10中的枢纽节点11相互直接连接。枢纽节点11与枢纽节点11之间可以通过专网进行互联。枢纽节点11承担对参与节点12进行信息聚合、寻址导航等功能。参与节点12可以是各类政务主体、行业主体、公司主体、机构主体等。直接连接到同一个枢纽节点11的参与节点12通过该枢纽节点11进行通信。直接连接到不同枢纽节点11的参与节点12通过它们各自直接连接的枢纽节点11进行通信。也就是说,参与节点12只与其直接连接的枢纽节点11直接通信,枢纽节点11之间可直接通信,而参与节点12之间需经由相应的枢纽节点11进行通信。

[0043] 在实践中,数联网中可能存在海量的子网10。单个子网10中可能存在海量的参与节点12。因此,数联网中的参与节点12的数量可能是非常庞大的。各个参与节点12之间可能进行数据交易,也可能共同执行计算任务。在一些应用场景下,进行数据交易的需求方(或者发起共同计算任务的发起方)可能需要对目标节点的数据质量进行评估以便确定是否与该目标节点进行数据交易(或者共同执行计算任务)。

[0044] 本公开提出了一种对目标节点进行数据质量评估的方法。目标节点可以是互联网中的一个节点。目标节点也可以是数联网中的一个节点。目标节点可以是数联网中的参与节点(即,数据源(用户)节点),也可以是数据集节点(即,单个数据集被看作为节点)。图2示出根据本公开的实施例的对目标节点进行数据质量评估的方法200的示意性流程图。该方法200由目标节点执行。

[0045] 在图2的框S202处,确定目标节点是否接收到需求方的数据质量评估请求。在这里,需求方指的是需要对目标节点的数据质量进行评估的节点。目标节点的数据质量是相对于需求方而言的。需求方不包括该目标节点本身。

[0046] 如果接收到需求方的数据质量评估请求(在框S202处为“是”),则在框S204处目标节点与需求方联合计算目标评估指标并根据目标评估指标来评估目标节点的数据质量。如

果没有接收到需求方的数据质量评估请求(在框S202处为“否”),则目标节点不执行数据质量评估的操作,过程继续保持在框S202处。

[0047] 在本公开的一些实施例中,目标节点拥有第一数据集。第一数据集包括多个第一数据。需求方拥有第二数据集。第二数据集包括多个第二数据。目标评估指标包括第一评估指标。第一评估指标是根据目标节点的第一数据集中的该多个第一数据的方差膨胀因子(Variance Inflation Factor,简称VIF)来确定的。其中,每个第一数据的方差膨胀因子与需求方的第二数据集相关。

[0048] 在本公开的一些实施例中,第一评估指标通过以下操作来确定:计算第一数据集中方差膨胀因子超过预设阈值的第一数据的数量;以及将所计算的数量与第一数据集中的第一数据的总数量的比值作为第一评估指标。该预设阈值可根据具体应用来设定。如果第一评估指标高于第一评估指标阈值,则认为目标节点的数据质量不达标。第一评估指标阈值可根据具体应用来设定。

[0049] 图3示出根据本公开的实施例的计算第一数据的方差膨胀因子的步骤的示意性流程图。

[0050] 在图3的框S301处,目标节点与需求方联合计算第一数据集与第二数据集之间的皮尔逊相关系数矩阵。在上下文中,用P来表示皮尔逊相关系数矩阵。图4示出皮尔逊相关系数矩阵的示例图。如图4所示,皮尔逊相关系数矩阵P包括:多个第一数据两两之间的第一皮尔逊相关系数(图4的框A1中的皮尔逊相关系数),多个第二数据两两之间的第二皮尔逊相关系数(图4的框A2中的皮尔逊相关系数),多个第一数据与多个第二数据两两之间的第三皮尔逊相关系数(图4的框A3中的皮尔逊相关系数)。

[0051] 在图4的示例中,假设第一数据集包括m个第一数据x1,……,xm,第二数据集包括n个第二数据y1,……,yn。pear(x1,x1)表示x1与x1之间的第一皮尔逊相关系数,pear(x1,xm)表示x1与xm之间的第一皮尔逊相关系数,以此类推。pear(y1,y1)表示y1与y1之间的第二皮尔逊相关系数,pear(y1,yn)表示y1与yn之间的第二皮尔逊相关系数,以此类推。pear(x1,y1)表示x1与y1之间的第三皮尔逊相关系数,pear(x1,yn)表示x1与yn之间的第三皮尔逊相关系数,以此类推。

[0052] 任一数据X与任一数据Y的皮尔逊相关系数r,在数据X与数据Y是样本数据的情况下,可根据下式计算:

[0053] $$r=\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{X_i-\overline{X}}{\sigma X}\right)\left(\frac{Y_i-\overline{Y}}{\sigma Y}\right) \qquad (1)$$

[0054] 其中,N表示数据X中的特征值的个数(数据Y中的特征值的个数也为N),X_i表示数据X的第i特征值,$\overline{X}$表示数据X的所有特征值的平均值,σX表示数据X的标准差,$\frac{X_i-\overline{X}}{\sigma X}$表示数据X的标准分数(z-score),$Y_i$表示数据Y的第i特征值,$\overline{Y}$表示数据Y的所有特征值的平均值,σY表示数据Y的标准差,$\frac{Y_i-\overline{Y}}{\sigma Y}$表示数据Y的标准分数。

[0055] 根据式(1),可通过将计算皮尔逊相关系数r的逻辑拆分成三步来获得r。第一步:先将数据X和数据Y分别标准分数化(也可称为z-score标准化)。标准分数化的公式为:

[0056] $$z = \frac{s - \mu}{\sigma} \qquad (2)$$

[0057] 其中,s表示原始特征值,μ表示s序列的平均值,σ表示s序列的标准差,z表示标准分数化后的s。

[0058] 第二步:将标准分数化后的数据X的N个特征值构成的列向量与标准分数化后的数据Y的N个特征值构成的列向量对应相乘,并对乘积结果进行求和。第三步:将求和结果除以N-1。

[0059] 本领域技术人员应理解皮尔逊相关系数r的分母通常是N-1。这是因为皮尔逊相关系数是用来衡量两个变量之间的线性关系强度的统计量,通常用于样本数据而不是总体数据(全部数据)。在样本数据中,分母通常使用N-1,以校正样本的自由度,以更准确地估计总体的相关性。这样做有助于避免因为样本大小较小而导致的相关系数估计偏差。

[0060] 在某些特定情况下,当需要计算整个总体的皮尔逊相关系数时,分母可以是N而不是N-1。这通常发生在总体数据的分析中,而不是样本数据的情况。当有总体的全部数据时,可以使用N作为分母,因为不需要通过样本来估计总体参数,不需要校正自由度。

[0061] 总而言之,皮尔逊相关系数r的分母是N还是N-1取决于正在处理的数据是总体数据还是样本数据。如果是总体数据,可以使用N;如果是样本数据,通常使用N-1来更准确地估计总体的相关性。

[0062] N和N-1的区别经常出现在方差(variance)和标准差(standard deviation)的计算中,当使用样本数据来估计总体数据时,如果在计算样本方差时使用N作为分母,那么会得到样本方差的有偏估计;而使用N-1作为分母可以得到无偏估计,这称为贝塞尔校正(Bessel's correction)。在统计学中,当想要从样本方差推断总体方差时,通常使用N-1作为分母来修正样本大小的偏差,这样的计算被称为“样本方差”。

[0063] 在下文中以皮尔逊相关系数r的分母为N-1来进行示例性说明。

[0064] 假设数据X为$[1.2,3.3,1.5,4.1,3.5]^T$,数据Y为$[2.1,6.4,1.9,3.6,2.3]^T$,则标准分数化后的数据X为$[-1.3182,0.5030,-1.0580,1.1968,0.6764]^T$,且标准分数化后的数据Y为$[-0.6910,1.8704,-0.8101,0.2025,-0.5718]^T$。第二步的执行结果为:(-1.3182×-0.6910)+(0.5030×1.8704)+(-1.0580×-0.8101)+(1.1968×0.2025)+(0.6764×-0.5718)=2.56435968。第三步的执行结果为:2.56435968/(5-1)=0.6411。

[0065] 在实际应用中,第一数据集中的第一数据的数量远小于第一数据中的特征值的数量,因此无法从皮尔逊相关系数r反推出第一数据。类似地,也无法从皮尔逊相关系数r反推出第二数据。

[0066] 类似上述过程,目标节点可根据式(1)在本地计算第一皮尔逊相关系数。需求方可根据式(1)在本地计算第二皮尔逊相关系数,然后将第二皮尔逊相关系数发送给目标节点。

[0067] 在本公开的一些实施例中,第i第一数据xi与第j第一数据xj之间的第一皮尔逊相关系数通过以下操作来计算:将第i第一数据xi中的每个特征值进行标准分数化以获得第i第一变换数据;将第j第一数据xj中的每个特征值进行标准分数化以获得第j第一变换数据;将第i第一变换数据中的每个特征值与第j第一变换数据中的对应特征值之积进行求和以获得第一和;响应于第一数据集是目标节点的待评估数据中的全部数据,将第一和除以N的商作为第i第一数据xi与第j第一数据xj之间的第一皮尔逊相关系数;以及响应于第一数

据集是目标节点的待评估数据中的样本数据,将第一和除以N-1的商作为第i第一数据xi与第j第一数据xj之间的第一皮尔逊相关系数。其中,N等于第一数据中的特征值的个数。i等于j或者i不等于j。

[0068] 在本公开的一些实施例中,第i第二数据yi与第j第二数据yj之间的第二皮尔逊相关系数通过以下操作来计算:将第i第二数据yi中的每个特征值进行标准分数化以获得第i第二变换数据;将第j第二数据yj中的每个特征值进行标准分数化以获得第j第二变换数据;将第i第二变换数据中的每个特征值与第j第二变换数据中的对应特征值之积进行求和以获得第二和;响应于第一数据集是目标节点的待评估数据中的全部数据,将第二和除以N的商作为第i第二数据yi与第j第二数据yj之间的第二皮尔逊相关系数;以及响应于第一数据集是目标节点的待评估数据中的样本数据,将第二和除以N-1的商作为第i第二数据yi与第j第二数据yj之间的第二皮尔逊相关系数。其中,N等于第二数据中的特征值的个数。i等于j或者i不等于j。

[0069] 第一数据与第二数据之间的第三皮尔逊相关系数需要目标节点与需求方联合计算。图5示出根据本公开的实施例的计算任一第一数据与任一第二数据之间的第三皮尔逊相关系数的步骤的示意性流程图。

[0070] 在框S502处,目标节点将第一数据中的每个特征值进行标准分数化以获得第一变换数据。目标节点可根据式(2)对第一数据中的每个特征值进行标准分数化。假设第一数据包括特征值s1、s2和s3,经过标准分数化后,第一变换数据包括特征值z1、z2和z3。

[0071] 在框S504处,目标节点对第一变换数据中的每个特征值进行同态加密以获得第一加密变换数据。在第一变换数据包括特征值z1、z2和z3的示例中,特征值z1、z2和z3被同态加密之后,可得到Enc(z1)、Enc(z2)和Enc(z3)。换句话说,第一加密变换数据包括Enc(z1)、Enc(z2)和Enc(z3)。

[0072] 在框S506处,目标节点向需求方发送第一加密变换数据。

[0073] 在框S508处,目标节点接收由需求方生成的特征和。其中,特征和通过将第一加密变换数据中的每个加密特征值与第二数据中的对应特征值之积进行求和来生成。假设第二数据包括val1、val2和val3,则将Enc(z1)、Enc(z2)和Enc(z3)与val1、val2和val3分别对应相乘,可得到nv1=Enc(z1)×val1,nv2=Enc(z2)×val2,nv3=Enc(z3)×val3。然后将nv1、nv2和nv3相加得到特征和enc_sum_val。

[0074] 在框S510处,目标节点对所接收的特征和进行同态解密以获得解密后的特征和。enc_sum_val可被解密以获得解密后的特征和dec(enc_sum_val)。dec(enc_sum_val)的值实际上就等于z1×val1+z2×val2+z3×val3。

[0075] 在框S512处,目标节点将解密后的特征和除以N-1或者N的商作为第三皮尔逊相关系数。在第一数据集是目标节点的待评估数据中的全部数据的情况下,目标节点将解密后的特征和除以N的商作为第三皮尔逊相关系数。即,第一数据与第二数据之间的第三皮尔逊相关系数等于dec(enc_sum_val)/N。在第一数据集是目标节点的待评估数据中的采样数据的情况下,目标节点将解密后的特征和除以N-1的商作为第三皮尔逊相关系数。即,第一数据与第二数据之间的第三皮尔逊相关系数等于dec(enc_sum_val)/(N-1)。其中,N等于第一数据中的特征值的个数。

[0076] 目标节点与需求方联合计算第一数据与第二数据之间的第三皮尔逊相关系数的

过程中,由于目标节点向需求方发送的是经同态加密的数据,因此,目标节点的第一数据不会被泄露给需求方。在实际应用中,第二数据的特征值的个数较大,因此,目标节点也无法从需求方发送的特征和中求解出第二数据。即,需求方的第二数据也不会被泄露给目标节点。这样,可以实现对目标节点的数据质量的安全评估。

[0077] 目标节点可以根据在本地计算的第一皮尔逊相关系数,从需求方获得的第二皮尔逊相关系数以及与需求方联合计算的第三皮尔逊相关系数,拼接得到第一数据集与第二数据集之间的皮尔逊相关系数矩阵P。

[0078] 回到图3,在框S302处根据皮尔逊相关系数矩阵来计算第一数据的方差膨胀因子。根据皮尔逊相关系数矩阵来计算第一数据的方差膨胀因子的步骤可包括框S304、S306和S308处的操作。

[0079] 在框S304处,确定皮尔逊相关系数矩阵是否是满秩矩阵。如果皮尔逊相关系数矩阵是满秩矩阵(在框S304处为“是”),则在框S306处根据下式(3)来计算第一数据的方差膨胀因子:

[0080] $$VIF_i = \frac{|M_{ii}|}{|P|} \qquad (3)$$

[0081] 其中,VIF_i表示第一数据集中的第i第一数据的方差膨胀因子,$|P|$表示皮尔逊相关系数矩阵的行列式,$|M_{ii}|$表示皮尔逊相关系数矩阵的第i行和第i列被去除之后的部分的行列式。

[0082] 在根据式(3)来计算方差膨胀因子VIF_i的过程中,可针对第i第一数据,从皮尔逊相关系数矩阵P中删除第i行和第i列,以获得余子式M_{ii}。然后,可计算皮尔逊相关系数矩阵P的行列式$|P|$和余子式M_{ii}的行列式$|M_{ii}|$。然后将$|P|$和$|M_{ii}|$代入式(3)以获得VIF_i。

[0083] 如果皮尔逊相关系数矩阵不是满秩矩阵(在框S304处为“否”),则在框S308处根据下式(4)来计算第一数据的方差膨胀因子:

[0084] $$VIF_i = \frac{1}{1 - R_i^2} \qquad (4)$$

[0085] 其中,VIF_i表示第一数据集中的第i第一数据的方差膨胀因子,R_i^2表示第i第一数据对第一数据集中的其他所有第一数据的多元线性回归的决定系数,其中,R_i^2根据下式(5)来计算:

[0086] $$\begin{cases} R_i^2 = b' R_{-i,-i}^{-1} b, & Rmin < R_i^2 < Rmax \\ R_i^2 = Rmax, & R_i^2 \geq Rmax \\ R_i^2 = Rmin, & R_i^2 \leq Rmin \end{cases} \qquad (5)$$

[0087] 其中,b表示皮尔逊相关系数矩阵的第i列向量的第i行被去除之后的列向量,b′表示b的转置向量,$R_{-i,-i}^{-1}$表示皮尔逊相关系数矩阵的第i行和第i列被去除之后的子矩阵的逆矩阵,Rmax表示针对R_i^2设置的最大值,Rmin表示针对R_i^2设置的最小值。Rmax和Rmin可根据实际应用来设置。

[0088] 在根据式(4)和(5)来计算方差膨胀因子VIF_i的过程中,可针对第i第一数据,从皮尔逊相关系数矩阵P中删除第i行和第i列,以获得余子式M_{ii}。然后对余子式M_{ii}求伪逆矩阵

以获得$R_{-i,-i}^{-1}$。针对第i第一数据，从皮尔逊相关系数矩阵P中提取第i第一数据与其它第一数据的皮尔逊相关系数以获得列向量b(例如，先获得P的第i列向量，然后从第i列向量中去除第i行)。将列向量b转置以获得b′。将$R_{-i,-i}^{-1}$、b和b′代入式(5)以获得R_i^2。将R_i^2代入式(4)以获得VIF_i。

[0089] 在上述实施例中，通过判断皮尔逊相关系数矩阵是否是满秩矩阵来选择对方差膨胀因子不同的计算方式，能够降低计算复杂度，从而实现更快的运行速度。

[0090] 在本公开的一些实施例中，目标评估指标还包括：第一数据集的信息价值。对目标节点的数据质量的评估可基于第一评估指标阈值和第一数据集的信息价值共同作出。其中，目标节点可通过以下操作(a)-(k)来计算第一数据集的信息价值。

[0091] 在操作(a)中，目标节点接收需求方发送的第二数据集的密态标签值和对应的唯一标识符，其中，密态标签值通过需求方对第二数据集的标签值进行同态加密来获得。

[0092] 在操作(b)中，目标节点根据所接收的唯一标识符来建立密态标签值与第一数据集的对应关系。

[0093] 在操作(c)中，目标节点根据第一数据集对应的分箱信息以及密态标签值得到每一分箱的密态模糊正样本数和密态模糊负样本数。其中，可用$E\left(r_1^i n_{正}^i\right)$来表示第i分箱的密态模糊正样本数，用$E\left(r_2^i n_{负}^i\right)$来表示第i分箱的密态模糊负样本数。在一个示例中，目标节点根据第一数据集对应的分箱信息以及密态标签值统计出密态正样本数$E\left(n_{正}^i\right)$和密态负样本数$E\left(n_{负}^i\right)$然后，针对每个分箱产生第一随机数r_1^i和第二随机数r_2^i。利用第一随机数r_1^i对密态正样本数$E\left(n_{正}^i\right)$进行混淆，得到密态模糊正样本数$E\left(r_1^i n_{正}^i\right)$；利用第二随机数$r_2^i$对密态负样本数$E\left(n_{负}^i\right)$进行混淆，得到密态模糊负样本数

[0094] $E\left(r_2^i n_{负}^i\right)$。

[0095] 在操作(d)中，目标节点向需求方发送每一分箱的密态模糊正样本数$E\left(r_1^i n_{正}^i\right)$和密态模糊负样本数$E\left(r_2^i n_{负}^i\right)$。

[0096] 在操作(e)中，目标节点接收需求方发送的每一分箱的模糊证据权重$\widetilde{woe}^i$、第一密态参数$E\left(C_1^i\right)$和第二密态参数$E\left(C_2^i\right)$。其中，每一分箱的模糊证据权重$\widetilde{woe}^i$、第一密态参数$E\left(C_1^i\right)$和第二密态参数$E\left(C_2^i\right)$是根据每一分箱的密态模糊正样本数$E\left(r_1^i n_{正}^i\right)$和密态模糊负样本数$E\left(r_2^i n_{负}^i\right)$生成的。在一个示例中，需求方对每一分箱的密态模糊正样本数$E\left(r_1^i n_{正}^i\right)$和密态模糊负样本数$E\left(r_2^i n_{负}^i\right)$进行解密可获得每一分箱的模糊正样本数$r_1^i n_{正}^i$和模糊负样本数$r_2^i n_{负}^i$。根据每一分箱的模糊正样本数$r_1^i n_{正}^i$、模糊负样本数$r_2^i n_{负}^i$、正样本

个数$n_{正}^{all}$和负样本个数$n_{负}^{all}$，得到所有分箱的模糊证据权重：$\widetilde{woe}^i = \log\left(\frac{n_{负}^{all}}{n_{正}^{all}} \times \frac{r_1^i n_{正}^i}{r_2^i n_{负}^i}\right)$。需求方根据模糊正样本数$r_1^i n_{正}^i$和正样本个数$n_{正}^{all}$得到第一中间参数$c_1^i = \frac{r_1^i n_{正}^i}{n_{正}^{all}}$，需求方根据模糊负样本数$r_1^i n_{负}^i$和负样本个数$n_{负}^{all}$得到第二中间参数$c_2^i = \frac{r_2^i n_{负}^i}{n_{负}^{all}}$。需求方对第一中间参数$c_1^i$进行加密得到第一密态参数$E(C_1^i)$。需求方对第二中间参数$c_2^i$进行加密得到第二密态参数$E(C_2^i)$。

[0097] 在操作(f)中，目标节点根据每一分箱的模糊证据权重$\widetilde{woe}^i$得到每一分箱的证据权重。在一个示例中，目标节点可根据第一随机数r_1^i和第二随机数r_2^i对模糊证据权重$\widetilde{woe}^i$进行解模糊操作以得到证据权重$woe^i = \widetilde{woe}^i - \log\left(\frac{r_1^i}{r_2^i}\right)$。

[0098] 在操作(g)中，目标节点根据每一分箱的第一密态参数$E(C_1^i)$和第二密态参数$E(C_2^i)$得到每一分箱的密态模糊权重系数$E(C_3^i)$。在一个示例中，目标节点对应每一分箱产生第三随机数r_3^i。然后，对第一密态参数$E(C_1^i)$和第二密态参数$E(C_2^i)$计算权重系数并混淆第三随机数，得到密态模糊权重系数：

[0099] $$E(C_3^i) = \frac{1}{r_1^i}E(C_1^i) - \frac{1}{r_2^i}E(C_2^i) + r_3^i \qquad (6)$$

[0100] 其中，i表示分箱数，$E(C_3^i)$表示密态模糊权重系数，$E(C_1^i)$表示第一密态参数，$E(C_2^i)$表示第二密态参数，r_3^i表示第三随机数。

[0101] 在操作(h)中，目标节点向需求方发送每一分箱的密态模糊权重系数

[0102] $E(C_3^i)$。

[0103] 在操作(i)中，目标节点接收需求方发送的每一分箱的模糊权重系数$\widetilde{coef}^i$。其中，每一分箱的模糊权重系数$\widetilde{coef}^i$是通过对每一分箱的密态模糊权重系数$E(C_3^i)$进行同态解密获得的。

[0104] 在操作(j)中，目标节点根据每一分箱的模糊权重系数得到每一分箱的权重系数。在一个示例中，目标节点通过将每一分箱的模糊权重系数$\widetilde{coef}^i$与对应的第三随机数r_3^i做差，得到每一分箱的权重系数$(\widetilde{coef}^i - r_3^i)$。

[0105] 在操作(k)中，目标节点对每一分箱的证据权重woe^i与对应的权重系数$(\widetilde{coef}^i - r_3^i)$之积进行求和以获得第一数据集的信息价值。

[0106] $$IV = \sum_{i=0}^{m} \left(\widetilde{coef}^{i} - r_3^i\right) \times woe^i \qquad (7)$$

[0107] 其中，IV表示第一数据集的信息价值，woe^i表示第i分箱的证据权重，$\left(\widetilde{coef}^{i} - r_3^i\right)$表示第i分箱的权重系数。

[0108] 在本公开的一些实施例中，可在更多维度上对数据质量进行评估。目标评估指标还可包括：对第一数据集的本地质量评估指标。其中，本地质量评估指标包括以下中的一个或多个：完整性、准确性、一致性、可用性、及时性、可理解性、异常值、可信度、关联性、数据唯一性、敏感性、数据规模、数据存储效率、数据安全性、数据历史变化程度、数据可操作性。

[0109] 完整性例如包括缺失值比例和完整性约束方面的评估。缺失值比例用于衡量数据集中缺失值的百分比。更低的缺失值比例通常表示更高的数据完整性。完整性约束用于定义和应用数据完整性约束，以确保数据满足特定标准和规则。

[0110] 准确性例如包括数据错误率和数据验证方面的评估。数据错误率指的是识别和记录数据中的错误或不准确信息的比例。数据验证指的是使用规则和验证方法来检测数据的准确性。

[0111] 一致性例如包括数据一致性检查和一致性规则方面的评估。数据一致性检查指的是比较相同数据在不同数据源中的一致性，确保数据一致。一致性规则指的是定义和应用一致性规则以确保数据在不同数据源之间保持一致。

[0112] 可用性例如包括数据可用性时间和数据访问权限方面的评估。数据可用性时间指的是测量数据可用于分析和决策的时间，包括数据更新的频率。数据访问权限用于确保只有授权用户可以访问数据，以维护数据的安全性。

[0113] 及时性例如包括数据更新频率和数据交付时间方面的评估。数据更新频率指的是衡量数据更新的频率，以确保数据保持最新。数据交付时间涉及测量数据可在需要时提供的速度。

[0114] 可理解性例如包括数据文档、数据标签和元数据方面的评估。数据文档评估数据是否提供详细的数据文档，以帮助用户理解数据的含义和结构。数据标签和元数据指的是使用元数据和数据标签来描述数据，使其更易于理解和使用。

[0115] 异常值指的是识别和记录数据中的异常值，以反映数据质量问题的迹象。

[0116] 可信度包括数据来源可信度。数据来源可信度衡量数据的来源的可信度和信誉度，确保数据来自可信赖的来源。

[0117] 关联性指的是数据关联度，用于评估数据之间的关联和连接，以确保数据在不同数据集之间保持一致。

[0118] 数据唯一性包括重复数据方面的评估。重复数据指的是检测和删除数据中的重复记录，以确保数据的唯一性。

[0119] 敏感性包括敏感数据处理方面的评估。敏感数据处理指的是确保对包含敏感信息的数据采取适当的保护措施，以保护隐私和遵守法规。

[0120] 数据规模包括数据规模适应性方面的评估。数据规模适应性用于确保数据质量度量方法和工具能够适应大规模数据处理需求。

[0121] 数据存储效率用于评估数据存储结构和性能，以确保数据可以高效地存储和检

索。

[0122] 数据安全性用于保护数据免受未经授权的访问和威胁,确保数据的机密性和完整性。

[0123] 数据历史变化程度指的是跟踪数据的历史变化,以便追溯和审计数据变更。

[0124] 数据可操作性用于确保数据易于操作和分析,以满足业务需求。

[0125] 这些维度可能根据不同情境和行业有所变化。选择合适的数据质量维度取决于具体数据质量管理目标和需求。在数据质量管理过程中,通常需要综合考虑多个维度,以确保数据在整个数据生命周期中保持高质量。

[0126] 图6示出根据本公开的实施例的对目标节点进行数据质量评估的装置600的示意性框图。该装置600位于目标节点上。如图6所示,该装置600可包括处理器610和存储有计算机程序的存储器620。当计算机程序由处理器610执行时,使得装置600可执行如图2所示的方法200的步骤。在一个示例中,装置600可以是计算机设备或云计算节点。响应于接收到需求方的数据质量评估请求,装置600可与需求方联合计算目标评估指标并根据目标评估指标来评估目标节点的数据质量。其中,目标评估指标包括:根据目标节点的第一数据集中的多个第一数据的方差膨胀因子来确定的第一评估指标。每个第一数据的方差膨胀因子与需求方的第二数据集相关。第二数据集包括多个第二数据。装置600可与需求方联合计算第一数据集与第二数据集之间的皮尔逊相关系数矩阵,并根据皮尔逊相关系数矩阵来计算第一数据的方差膨胀因子。其中,皮尔逊相关系数矩阵包括:多个第一数据两两之间的第一皮尔逊相关系数,多个第二数据两两之间的第二皮尔逊相关系数,多个第一数据与多个第二数据两两之间的第三皮尔逊相关系数。其中,第一皮尔逊相关系数由目标节点在本地计算。第二皮尔逊相关系数由需求方在本地计算之后发送给目标节点。装置600可将第一数据中的每个特征值进行标准分数化以获得第一变换数据。装置600可对第一变换数据中的每个特征值进行同态加密以获得第一加密变换数据。装置600可向需求方发送第一加密变换数据。装置600可接收由需求方生成的特征和。其中,特征和通过将第一加密变换数据中的每个加密特征值与第二数据中的对应特征值之积进行求和来生成。装置600可对所接收的特征和进行同态解密以获得解密后的特征和。响应于第一数据集是目标节点的待评估数据中的全部数据,装置600可将解密后的特征和除以N的商作为第三皮尔逊相关系数。响应于第一数据集是目标节点的待评估数据中的样本数据,装置600可将解密后的特征和除以N-1的商作为第三皮尔逊相关系数。其中,N等于第一数据中的特征值的个数。

[0127] 在本公开的实施例中,处理器610可以是例如中央处理单元(CPU)、微处理器、数字信号处理器(DSP)、基于多核的处理器架构的处理器等。存储器620可以是使用数据存储技术实现的任何类型的存储器,包括但不限于随机存取存储器、只读存储器、基于半导体的存储器、闪存、磁盘存储器等。

[0128] 此外,在本公开的实施例中,装置600也可包括输入设备630,例如键盘、鼠标等,用于输入第一数据集。另外,装置600还可包括输出设备640,例如显示器等,用于输出数据质量评估结果。

[0129] 在本公开的其它实施例中,还提供了一种存储有计算机程序的计算机可读存储介质,其中,计算机程序在由处理器执行时能够实现如图2、图3和图5所示的方法的步骤。

[0130] 综上所述,根据本公开的实施例的对目标节点进行数据质量评估的方法和装置可

以实现目标节点与需求方联合对目标节点的数据质量进行安全评估,并且不会在评估过程中泄漏任何一方的原始数据信息。进一步地,本公开的实施例还考虑了运算量的问题,通过判断皮尔逊相关系数矩阵是否是满秩矩阵来选择对第一数据的方差膨胀因子不同的计算方式,能够降低计算复杂度,从而实现更快的运行速度。本公开的实施例还能够从多个维度对数据质量进行评估,适应更多应用场景。

[0131] 附图中的流程图和框图显示了根据本公开的多个实施例的装置和方法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。

[0132] 除非上下文中另外明确地指出,否则在本文和所附权利要求中所使用的词语的单数形式包括复数,反之亦然。因而,当提及单数时,通常包括相应术语的复数。相似地,措辞“包含”和“包括”将解释为包含在内而不是独占性地。同样地,术语“包括”和“或”应当解释为包括在内的,除非本文中明确禁止这样的解释。在本文中使用术语“示例”之处,特别是当其位于一组术语之后时,所述“示例”仅仅是示例性的和阐述性的,且不应当被认为是独占性的或广泛性的。

[0133] 适应性的进一步的方面和范围从本文中提供的描述变得明显。应当理解,本申请的各个方面可以单独或者与一个或多个其它方面组合实施。还应当理解,本文中的描述和特定实施例旨在仅说明的目的并不旨在限制本申请的范围。

[0134] 以上对本公开的若干实施例进行了详细描述,但显然,本领域技术人员可以在不脱离本公开的精神和范围的情况下对本公开的实施例进行各种修改和变型。本公开的保护范围由所附的权利要求限定。

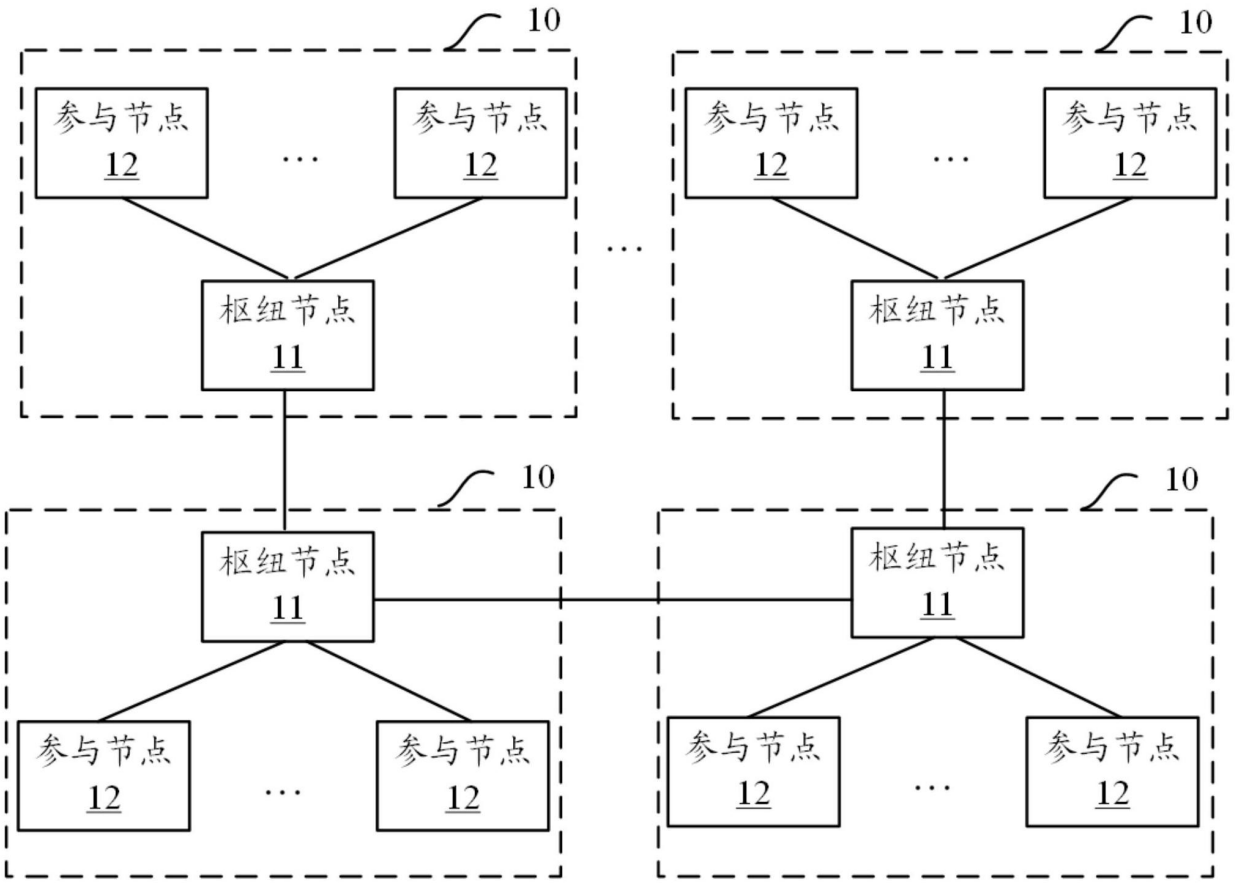

10
参与节点
12
…
参与节点
12
枢纽节点
11
10
参与节点
12
…
参与节点
12
枢纽节点
11
10
枢纽节点
11
10
枢纽节点
11
参与节点
12
…
参与节点
12
参与节点
12
…
参与节点
12

图1

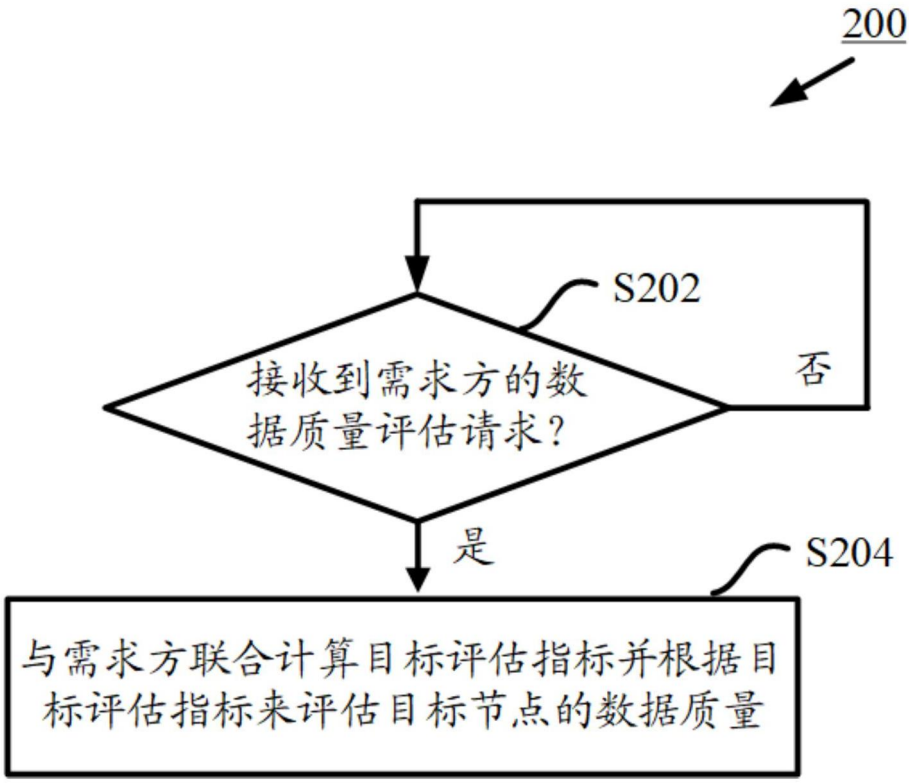

200
S202
接收到需求方的数据质量评估请求？
否
是
S204
与需求方联合计算目标评估指标并根据目标评估指标来评估目标节点的数据质量

图2

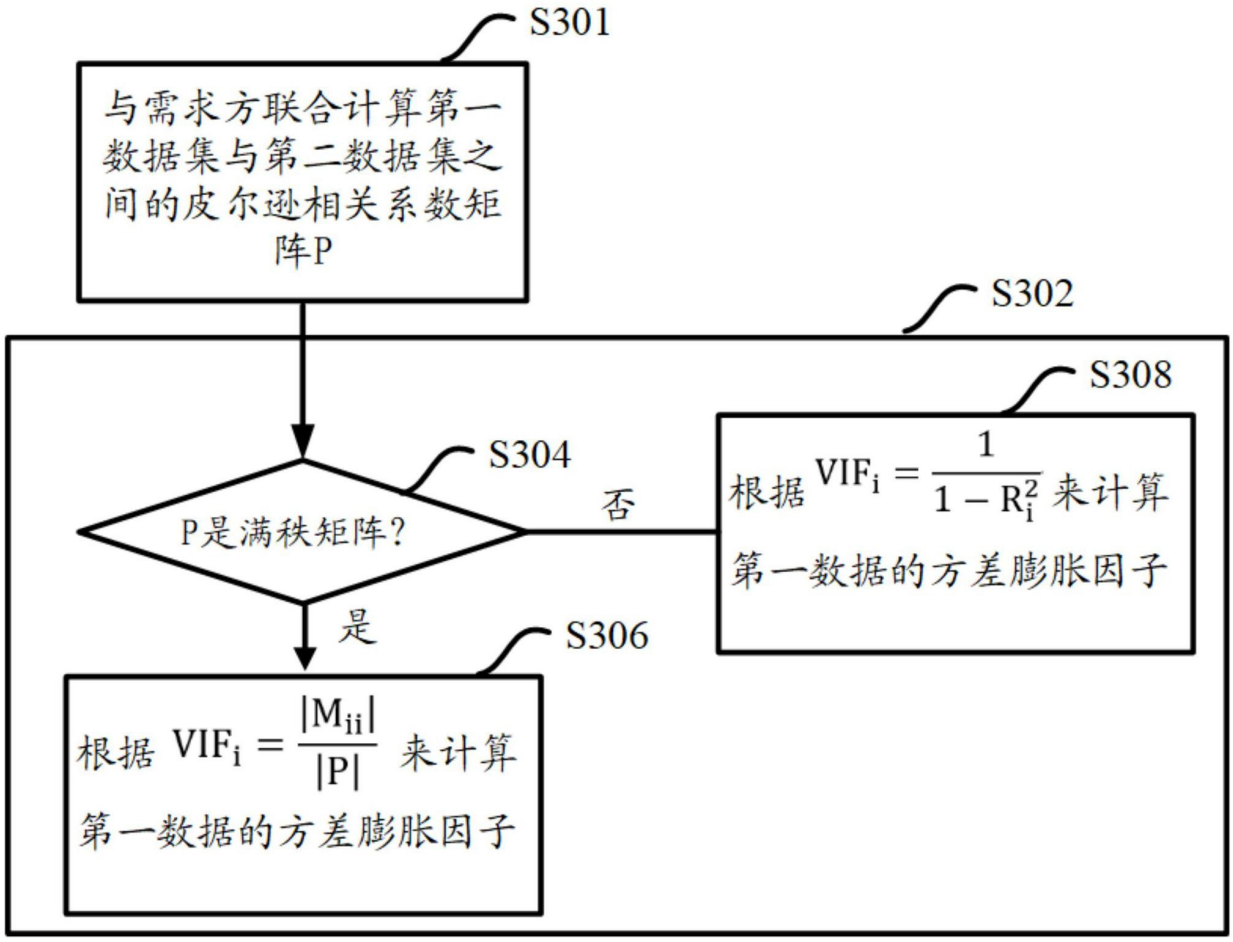

S301
与需求方联合计算第一数据集与第二数据集之间的皮尔逊相关系数矩阵P
S302
S308
根据 $VIF_i = \frac{1}{1-R_i^2}$ 来计算第一数据的方差膨胀因子
S304
P是满秩矩阵?
否
是
S306
根据 $VIF_i = \frac{|M_{ii}|}{|P|}$ 来计算第一数据的方差膨胀因子

图3

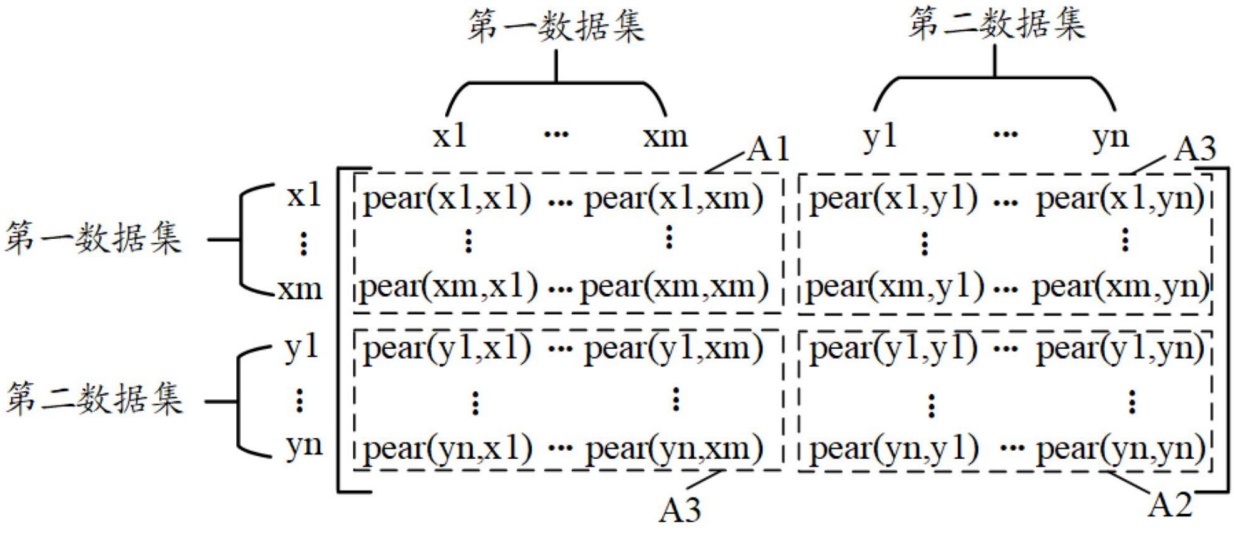

第一数据集
第二数据集
x1 … xm
y1 … yn
A1
A3
第一数据集
第二数据集
x1
xm
y1
yn
pear(x1,x1) … pear(x1,xm)
pear(xm,x1) … pear(xm,xm)
pear(x1,y1) … pear(x1,yn)
pear(xm,y1) … pear(xm,yn)
pear(y1,x1) … pear(y1,xm)
pear(yn,x1) … pear(yn,xm)
pear(y1,y1) … pear(y1,yn)
pear(yn,y1) … pear(yn,yn)
A3
A2

图4

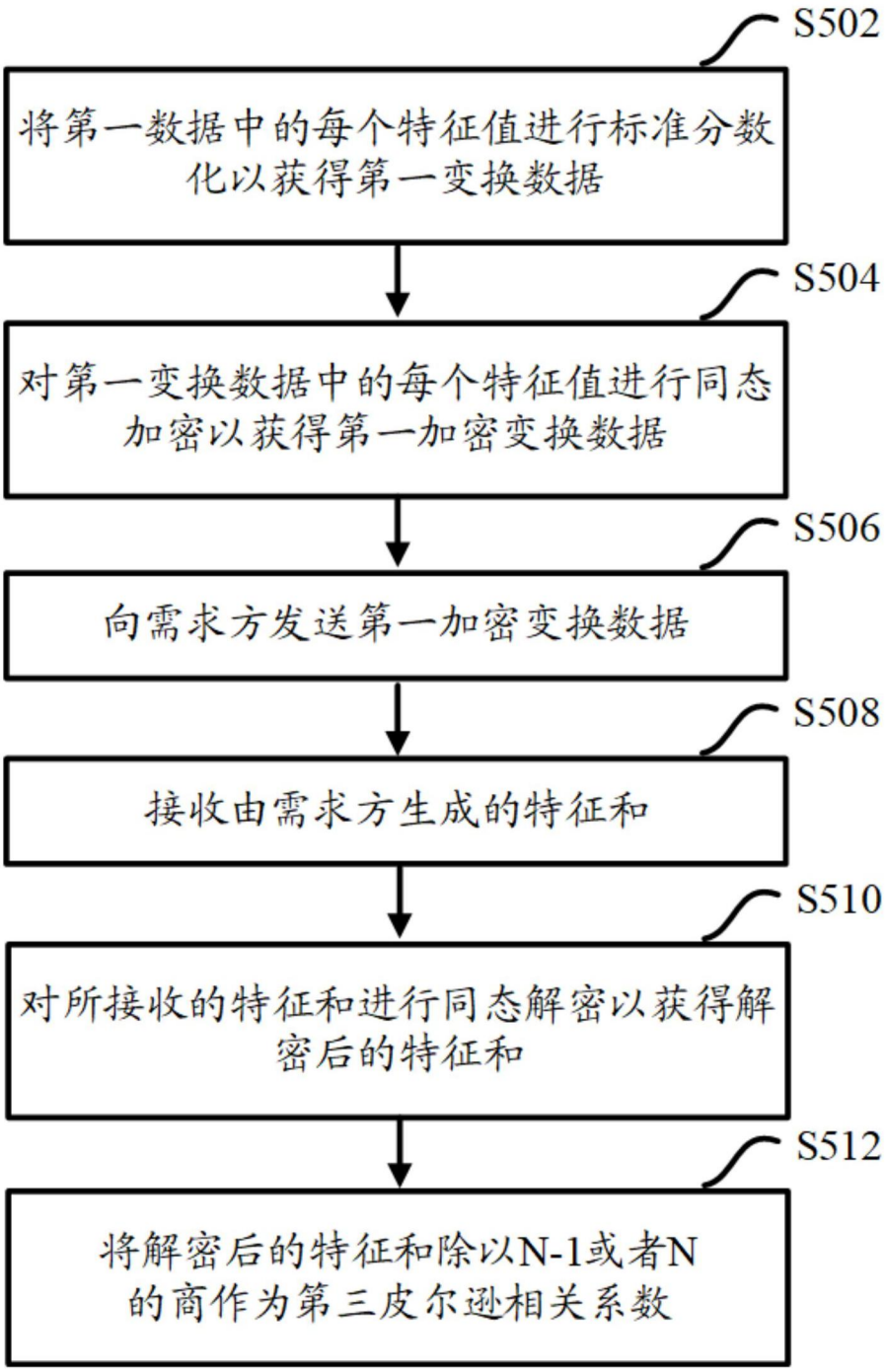

S502
将第一数据中的每个特征值进行标准分数化以获得第一变换数据
S504
对第一变换数据中的每个特征值进行同态加密以获得第一加密变换数据
S506
向需求方发送第一加密变换数据
S508
接收由需求方生成的特征和
S510
对所接收的特征和进行同态解密以获得解密后的特征和
S512
将解密后的特征和除以N-1或者N的商作为第三皮尔逊相关系数

图5

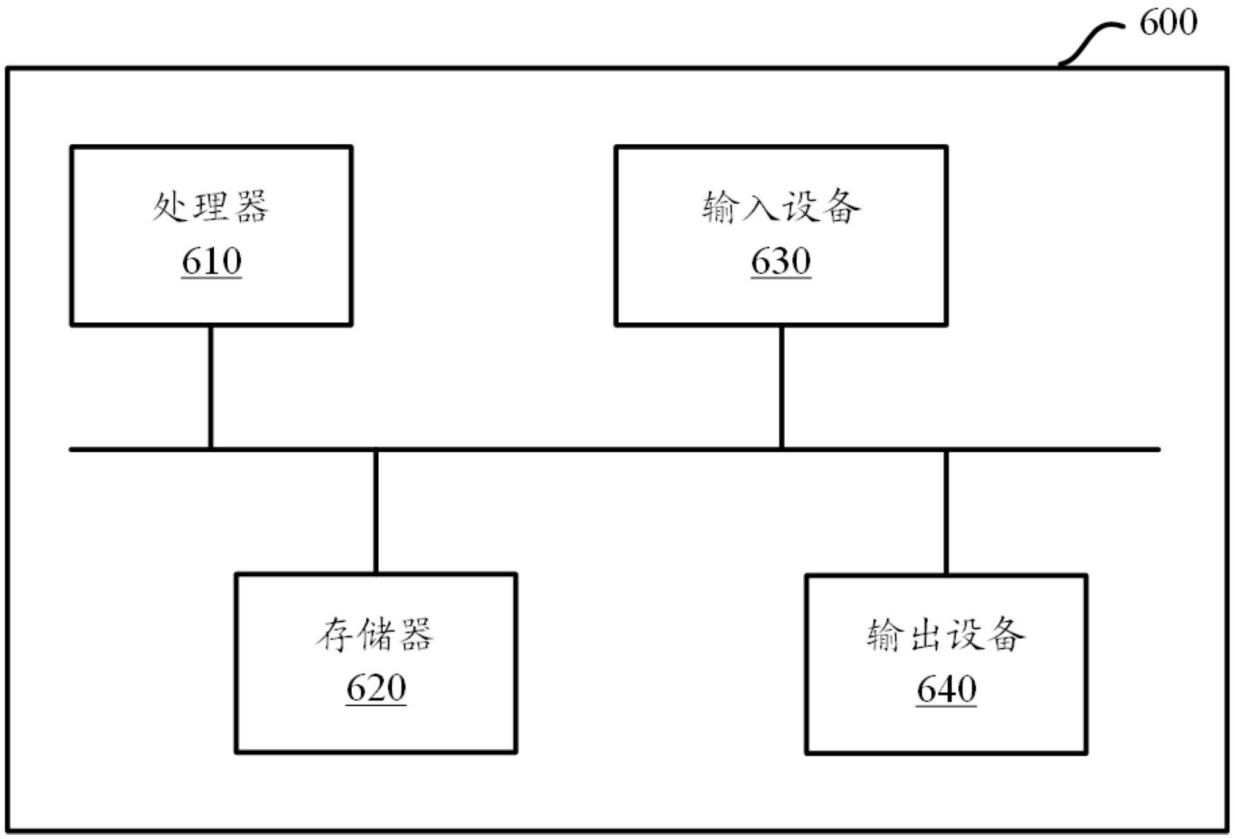


图6