

确定数据异常的方法及装置

技术领域

[01] 本说明书一个或多个实施例涉及计算机技术领域，尤其涉及确定数据异常的方法和装置。

5 背景技术

[02] 随着互联网的升级，不同平台之间的数据交互越来越频繁。例如，用户向银行或某个金融平台提交贷款请求，银行或金融平台会将用户的请求数据发送到综合计算平台进行全面评估，来决定是否批准该用户的贷款请求，以及核准的贷款金额。然而，各个平台为了安全性和自身用户保密性的考虑，在将数据发送给其他平台处理之前，往往会对数据进行一些隐私保护处理。这样处理之后的数据会失去其业务含义。接收平台在接收到这些数据之后，很难根据业务规则对数据中是否存在异常进行判断。例如，数据传送过程中是否被攻击，是否被篡改，用户群体是否发生了偏移等。

[03] 即使是在同一计算平台中，数据的处理往往要经过业务链中的多个环节。数据也有可能在某个业务环节，或者不同业务环节之间的传输过程中出现异常，例如系统被攻击、模型出现异常等等。

[04] 另一方面，在大数据背景下，数据量指数增长，而业务规则又是不断变化难以穷尽的，因此仅通过业务规则来发现数据异常，工作量巨大而不够全面。

[05] 因此，需要更有效的方式，对数据的异常进行判断和预警。

发明内容

[06] 本说明书一个或多个实施例描述了一种方法和装置，可以不依赖于数据的业务含义，对数据的异常进行判断和预警。

[07] 根据第一方面，提供了一种确定数据异常的方法，包括：获取预定时间段内的多个数据包，所述多个数据包具有相同数据结构；获取具有所述相同数据结构的历史数据的历史分布；将所述多个数据包与所述历史分布进行比对；根据比对结果，确定是否存在数据异常。

[08] 根据一种实施方式，将多个数据包与所述历史分布进行比对包括：通过将所述多个

数据包代入所述历史分布，获取所述多个数据包在所述历史分布中的多个分布状态参数；将所述多个分布状态参数与预定的与分布状态相关的阈值进行比对，确定超出所述阈值的数据包的数目。相应地，所述根据比对结果，确定是否存在数据异常包括：根据所述超出所述阈值的数据包的数目，确定是否存在数据异常。

- 5 [09]根据一种实施方式，将所述多个数据包与所述历史分布进行比对包括：确定所述多个数据包的数据分布状态作为当前分布；将所述当前分布与所述历史分布进行比对。

[10]在一种实施例，将所述当前分布与所述历史分布进行比对包括：确定所述当前分布的分布中心；获取所述历史分布的分布中心；确定所述当前分布的分布中心与历史分布的分布中心之间的偏移。相应地，所述根据比对结果，确定是否存在数据异常包括：

- 10 响应于所述偏移超出预定偏移阈值，确定存在数据异常。

[11]在另一实施例中，将所述当前分布与所述历史分布进行比对包括：确定随机抽取的数据包在所述当前分布中的分布状态参数，即第一参数；确定该随机抽取的数据包在所述历史分布中的分布状态参数，即第二参数；确定所述第一参数和第二参数的差值。相应地，所述根据比对结果，确定是否存在数据异常包括：响应于所述差值超出预定差值

- 15 阈值，确定存在数据异常。

[12]根据一种实施方式，所述历史分布是采用混合高斯模型对所述历史数据进行处理得到的历史概率分布；相应地，当前分布体现为采用混合高斯模型对当前多个数据包处理得到的当前概率分布；上述分布状态参数可以体现为概率值；分布中心体现为概率分布曲线的峰值位置。

- 20 [13]根据另一种实施方式，所述历史分布是针对所述历史数据采用聚类算法获得的历史聚类分布；相应地，当前分布体现为采用同样的聚类算法对当前多个数据包处理得到的当前聚类分布；上述分布状态参数可以体现为聚类分布空间中数据包的位置和对应聚类中心的距离；分布中心体现为聚类分布的中心位置。

- 25 [14]根据第二方面，提供一种确定数据异常的装置，包括：数据包获取单元，配置为获取预定时间段内的多个数据包，所述多个数据包具有相同数据结构；历史获取单元，配置为获取具有所述相同数据结构的历史数据的历史分布；比对单元，配置为将所述多个数据包与所述历史分布进行比对；确定单元，配置为根据比对结果，确定是否存在数据异常。

[15]根据第三方面，提供了一种计算机可读存储介质，其上存储有计算机程序，当所述

计算机程序在计算机中执行时，令计算机执行第一方面的方法。

[16]根据第四方面，提供了一种计算设备，包括存储器和处理器，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现上述第一方面的方法。

- 5 [17]通过本说明书实施例提供的方法及装置，将当前获得的数据包与基于历史数据统计获得的历史分布进行比对，根据比对结果确定当前数据包是否存在数据异常。如此，可以不依赖于数据的业务含义，而有效地对数据异常进行判断和预警。

附图说明

- 10 [18]为了更清楚地说明本发明实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其它的附图。

[19]图 1 示出本说明书披露的一个实施例的示意图；

[20]图 2 示出根据一个实施例的方法的流程图；

[21]图 3 示出根据一个实施例的比对和判断过程流程图；

- 15 [22]图 4 示出根据另一个实施例的比对和判断过程流程图；

[23]图 5 示出比对峰值位置的示意图；

[24]图 6 示出根据又一个实施例的比对和判断过程流程图；

[25]图 7 示出比对概率值的示意图；

[26]图 8 示出根据一个实施例的确定装置的示意框图。

20 具体实施方式

[27]下面结合附图，对本说明书提供的方案进行描述。

- 25 [28]图 1 为本说明书披露的一个实施例的示意图。在图 1 中，计算平台、例如支付宝服务器获取预定时间段内的多个数据包，例如用户请求贷款的信贷请求数据包。这些数据包具有相同的数据结构，例如具有相同的字段，字段内容可以包括用户年龄、性别、收入、贷款额等等。另一方面，计算平台还获取具有以上相同数据结构的历史数据的历史分布，这些历史数据可以由之前一段较长时间段内接收的大量同类数据包构成。接着，

计算平台将多个数据包与所述历史分布进行比对, 根据比对结果确定是否存在数据异常。如果不存在数据异常, 计算平台可以继续处理这些数据, 或者将这些数据发送到下一业务环节。如果确定存在数据异常, 则可以启动预警, 通知相关人员分析数据异常的原因, 触发相关解决方案。下面描述确定数据异常的具体实施过程。

- 5 [29] 图 2 示出根据一个实施例的方法流程图。该方法的执行主体可以是任何具有计算能力和处理能力的计算平台, 例如服务器。如图 2 所示, 该方法包括: 步骤 21, 获取预定时间段内的多个数据包, 所述多个数据包具有相同数据结构; 步骤 22, 获取具有所述相同数据结构的历史数据的历史分布; 步骤 23, 将所述多个数据包与所述历史分布进行比对; 步骤 24, 根据比对结果确定是否存在数据异常。下面结合具体例子描述以上各个步骤的执行方式。

10 [30] 首先在步骤 21, 获取预定时间段内的多个数据包。

- [31] 在一个实施例中, 所述多个数据包是从外部机构接收到的数据包, 例如支付宝服务器从银行或金融机构接收到的用户的信贷请求数据包。此时, 可以通过后续的步骤判断外部机构传输过来的数据包是否存在数据异常。这样的数据异常有可能是因为数据传送过程中被攻击、被篡改, 或者用户群体发生偏移等原因所致。

- [32] 在另一实施例中, 所述多个数据包是在数据处理的业务链中, 某个业务环节产生的数据包。例如, 对于从外部机构接收到的信贷请求数据包, 需要对其进行多个业务环节的处理, 包括字段解析、降维、模型运算等等。此时, 可以对任意业务环节中产生的数据包进行后续分析。通过后续分析, 可以判断该业务环节的数据包是否存在数据异常。
- 20 某个业务环节的数据包存在数据异常有可能是由于业务环节之间的数据传输出现问题, 或者该业务环节的系统或模型出现问题所致。因此, 对中间业务环节的数据包进行异常判断, 还可以帮助确定系统或模型是否存在异常。

- [33] 可以理解, 以上所述的多个数据包具有相同的数据结构。更具体而言, 每个数据包可以具有相同的字段。在一个例子中, 上述数据包为信贷请求数据包, 此时每个数据包的字段可以包括用户年龄、性别、收入、贷款额等等。在另一例子中, 上述数据包为用户操作记录, 用于分析用户的行为模式。此时, 每个数据包的字段可以包括用户 ID、操作行为、操作对象、操作时间等等。

- 25 [34] 在一个实施例中, 在步骤 21 中获取的多个数据包为距离当前时刻预定时间段内的数据包。上述预定时间段可被设置为较短的时间段, 从而使得获得的多个数据包为最近产

生的数据包。例如，在数据量较大的情况下，上述预定时间段可以预设为 0.5s 或 1s；在数据量相对较小的情况下，上述预定时间段可以预设为 30s 或者 1min。获取的数据包的数目也可以根据业务需要进行选择。在一个实施例中，获取上述预定时间段内的全部数据包。在另一实施例中，获取上述预定时间段内预定数目的数据包。

- 5 [35]在获得了上述具有相同结构的多个数据包的基础上，另一方面，在步骤 22，获取历史数据的历史分布。可以理解，上述历史数据与步骤 21 中获得的数据包具有相同的数
据结构。并且，一般地，历史数据包括之前一段较长时间段内、例如一周或一个月等接
收的大量同类数据包。在一个实施例中，上述历史数据与步骤 21 中获得的数据包具有
10 相同的数据来源。例如，在步骤 21 中的数据包是来自外部机构的信贷请求数据包的情
况下，对应地，上述历史数据也是来自外部机构的、之前一段较长时间段内接收到的信
贷请求数据包。在步骤 21 中的数据包是中间业务环节产生的数据包的情况下，对应地，
上述历史数据是之前一段较长时间段内、同样的业务环节产生的数据包。

[36]通过对历史数据进行统计、运算，可以获得历史数据的历史分布。

- [37]接下来，在步骤 23，将当前获得的多个数据包与历史分布进行比对；在步骤 24，根
15 据比对结果，确定当前数据包是否存在数据异常。

[38]步骤 23 中将当前的数据包与历史分布进行比对有多种实施方式，对应地，步骤 24
根据不同比对方式产生的比对结果进行判断。

- [39]在一个实施方式中，采用代入的方式进行比对。图 3 示出根据一个实施例的比对和
判断过程流程图。可以理解，图 3 所示的流程是图 2 中步骤 23 和 24 的子步骤。如图 3
20 所示，该流程包括以下步骤。在步骤 31，将多个数据包代入上述历史分布，从而获取所
述多个数据包各自在历史分布中的分布状态参数，得到多个分布状态参数；在步骤 32，
将上述多个分布状态参数与预定的与分布状态相关的阈值进行比对，确定超出阈值的数
据包的数目；接着，在步骤 33，根据所述超出阈值的数据包的数目，确定是否存在数据
异常。

- 25 [40]以上的流程是将当前获取的多个数据包代入到历史分布中，确定这多个数据包在历
史分布中的分布状态，如果出现大量（超出阈值）的异常状态，则认为当前存在数据异
常。

[41]如上所述，历史分布是对历史数据进行统计、运算而获得的数据分布状态。可以采
用不同算法，基于历史数据获得历史分布。下面结合高斯混合模型和聚类算法的例子，

描述在不同形式的历史分布的情况下，图 3 流程的具体执行过程。

[42] 在一个实施例中，上述历史分布是采用混合高斯模型（Gaussian Mixture Model, GMM）对历史数据进行处理得到的历史概率分布。可以理解，混合高斯模型 GMM 是用高斯概率密度函数精确地量化工事物的模型，它可以将一个事物分解为若干的基于高斯概率密度函数的形式。高斯概率密度函数通常表现为正态分布曲线的形式。如前所述，
5 不管是当前获得的各个数据包，还是历史数据，均具有相同的数据结构，该数据结构包含若干字段。在一个例子中，可以采用混合高斯模型，以数据字段为维度进行分解，从而确定出混合高斯概率密度函数作为历史概率分布。

[43] 在历史分布体现为高斯概率分布的情况下，图 3 的流程具体可以如下执行。在步骤
10 31，将当前获取的多个数据包代入历史概率分布，获取多个数据包分别在所述历史概率分布中的概率值，得到多个概率值。也就是说，在历史分布为高斯概率分布的情况下，分布状态参数体现为分布概率值。假定获取了 N 个数据包，那么在步骤 31，可以确定这 N 个数据包在历史概率分布中的概率值分别为 $p_1, p_2, \dots, p_N$ 。

[44] 在步骤 32，将确定的多个概率值与预定的概率阈值进行比对，确定概率值小于所述
15 概率阈值的数据包的数目，即第一数目。一般地，该预定的概率阈值为一较小的概率值 p_0 。概率阈值 p_0 的具体大小可以根据业务需要进行设置。在该步骤中，即确定概率值 $p_1, p_2, \dots, p_N$ 中小于概率阈值 p_0 的数目 $M1$ ($M1 \leq N$)。也就是说，确定当前数据包中有多少数据包为“小概率”数据包。

[45] 接着，在步骤 33，根据所述第一数目 $M1$ 确定是否存在数据异常。更具体地，在一
20 个例子中，如果第一数目 $M1$ 超出预定数目阈值 $M0$ ，则确定存在数据异常。也就是说，一旦当前获取的 N 个数据包中有超过 $M0$ 个数据包为“小概率”数据包，则认为当前存在异常。在另一例子中，如果第一数目 $M1$ 与所述多个数据包的数目 N 的比例 $M1/N$ 超出预定比例阈值，则确定存在数据异常。也就是说，一旦 N 个数据包有超过预定比例的数据包为“小概率”数据包，则认为当前存在数据异常。

25 [46] 如此，确定当前获取的数据包中“小概率”数据包的数目或比例，当该数目或比例超出一定阈值时，就可以认定，当前存在数据异常。

[47] 下面描述历史分布为聚类分布的情况下，图 3 的执行方式。可以理解，聚类算法是对多个样本进行相似性分析和数据统计的算法。对于具有相同数据结构的数据包而言，可以以数据包中的字段为维度，将每个数据包映射为多维空间中的点，然后采用聚类算

法将多个数据包进行划分和聚类。在聚类过程中,可以采取基于层次的聚类算法,例如 BIRCH 算法、CURE 算法等,或者采取基于密度的聚类算法,例如 DBSCAN 算法、OPTICS 算法等等,以及其他可能的聚类算法。

[48]在历史分布体现为聚类分布的情况下,图 3 的流程具体可以如下执行。在步骤 31,
5 确定当前获取的多个数据包各自在历史聚类分布的分布空间中的位置,得到多个位置,并确定所述多个位置与所述历史聚类分布的对应的聚类中心位置之间的距离,得到多个距离。如前所述,每个数据包可以映射为以字段为维度构成的多维空间中的点,该多维空间也就是历史聚类分布的分布空间。相应地,可以通过上述映射确定出当前获取的 N 个数据包在该历史聚类分布的空间中的位置 $P_1, P_2, \dots, P_N$ 。进一步地,还可以确定历史
10 聚类分布中聚类中心的位置,然后确定出各个位置 $P_1, P_2, \dots, P_N$ 与历史聚类分布对应的聚类中心位置之间的距离 $D_1, D_2, \dots, D_N$ 。可以理解,上述距离为多维空间中多维向量之间的距离。

[49]接着,在步骤 32,将所述多个距离与预设距离阈值进行比对,确定超出所述预设距离阈值的数目,即第二数目。一般地,聚类中心位于数据分布最为密集的区域,与聚类
15 中心的距离越远,对应数据包的出现频率或概率越低。因此,可以将上述预定距离阈值设置为一较大的距离 D_0 。距离阈值 D_0 的具体大小可以根据业务需要进行设置。相应地,在该步骤中,即确定距离 $D_1, D_2, \dots, D_N$ 中大于距离阈值 D_0 的数目 M_2 ($M_2 \leq N$)。也就是说,确定当前获取的数据包中有多少数据包为远离聚类中心的数据包。

[50]然后,在步骤 33,根据上述第二数目确定是否存在数据异常。具体地,在一个例子
20 中,如果所述第二数目 M_2 超出预定数目阈值 M_0 ,则认为存在数据异常。也就是说,一旦当前获取的 N 个数据包中有超过 M_0 个数据包远离聚类中心,则认为当前存在数据异常。在另一例子中,如果第二数目 M_2 与所述多个数据包的数目 N 的比例 M_2/N 超出预定比例阈值,则确定存在数据异常。也就是说,一旦当前获取的 N 个数据包有超过预定比例的数据包远离聚类中心,则认为当前存在数据异常。

25 [51]如此,确定当前获取的数据包中远离聚类中心的数据包的数目或比例,当该数目或比例超出一定阈值时,就可以认定,当前存在数据异常。

[52]以上结合混合高斯分布和聚类分布的例子描述了将当前数据包代入历史分布进行比对的实施方式。在当前获取的数据包的数目较多的情况下,例如超过一定数目(比如 200
30 个),还可以进一步确定当前获取的多个数据包的分布状态作为当前分布,将当前分布与历史分布进行比对,从而确定数据异常。

[53]具体地，图 4 示出根据另一个实施例的比对和判断过程流程图。可以理解，图 4 所示的流程是图 2 中步骤 23 和 24 的子步骤。如图 4 所示，在步骤 41，确定当前获取的多个数据包的数据分布状态作为当前分布；在步骤 42，确定当前分布的分布中心；在步骤 43，获取历史分布的分布中心；在步骤 44，确定当前分布的分布中心与历史分布的分布中心之间的偏移；在步骤 45，判断上述偏移是否超出预定偏移阈值，在超出预定偏移阈值的情况下，在步骤 46，确定存在数据异常。

[54]以上的流程是将当前分布的分布中心与历史分布的分布中心进行比对，当两者之间出现超过预定阈值的偏移时，则认为当前存在数据异常。下面仍然结合高斯混合模型和聚类算法的例子，描述在不同形式的历史分布的情况下，图 4 流程的具体执行过程。

10 [55]在一个实施例中，历史分布是采用混合高斯模型对历史数据进行处理得到的历史概率分布。相应地，在步骤 41，对当前获取的多个数据包采取同样的混合高斯模型进行处理，得到当前数据包的混合高斯分布作为当前概率分布。接着，在步骤 42，确定第一峰值位置 P_1 ，该第一峰值位置为所述当前概率分布对应的曲线的峰值位置。在步骤 43，获取第二峰值位置 P_2 ，所述第二峰值位置为历史概率分布对应的曲线的峰值位置。换而言之，在高斯概率分布的情况下，分布中心体现为概率分布曲线的峰值位置。

[56]然后，在步骤 44，确定第一峰值位置 P_1 和第二峰值位置 P_2 之间的位置偏移 ΔP 。并且，在步骤 45，判断所述位置偏移 ΔP 是否超出预定偏移阈值，如果位置偏移超出预定阈值，则确定当前存在数据异常。

20 [57]图 5 示出以上比对峰值位置的示意图。在图 5 中， V_1 为示意性示出当前概率分布的曲线， V_2 为示意性示出历史概率分布的曲线。可以看到，曲线 V_1 的峰值位置示意性示出为 P_1 ，曲线 V_2 的峰值位置示意性示出为 P_2 。如果峰值位置 P_1 和 P_2 之间的偏移 ΔP 大于预设阈值，那么可以认为，当前概率分布相比于历史概率分布发生了较大变化，因此当前有可能存在数据异常。

25 [58]可以理解的是，由于数据包中可能包含更多字段，相应地，混合高斯模型可能基于更多维度进行数据统计，实际的概率分布曲线图是立体多维的曲线图。图 5 仅仅是一种简化的示意。

[59]以上描述了在高斯分布的情况下，图 4 所示的比对分布中心的执行方式。在历史分布为采用聚类算法获得的聚类分布的情况下，可以类似地执行图 4 的流程。具体而言，在步骤 41，采用相同的聚类算法，确定当前获取的多个数据包的聚类分布作为当前聚类

分布。在步骤 42，确定第一中心位置，所述第一中心位置为所述当前聚类分布对应的聚类的中心位置；在步骤 43，获取第二中心位置，所述第二中心位置为所述历史聚类分布对应的聚类的中心位置。换言之，在聚类分布的情况下，分布中心体现为聚类分布的中心位置。可以理解，尽管不同的聚类算法可以将数据聚类为多个不同的簇或聚类，但是总是可以通过再次聚类或求重心的方式确定出整个聚类分布的中心。如此，分别确定出上述的第一中心位置和第二中心位置。

[60]接着，在步骤 44，确定上述第一中心位置和第二中心位置之间的距离 D 。可以理解，该距离可以是多维空间中多维点或多维向量之间的距离。

[61]然后，在步骤 45，判断上述距离 D 是否超出预定距离阈值。如果 D 超出距离阈值，则确定存在数据异常。

[62]在以上方式中，比对当前聚类分布的中心位置和历史聚类分布的中心位置，如果聚类分布的中心发生了较大偏移，则认为当前存在数据异常。

[63]除了比对分布中心之外，还可以随机抽取一个样本，比对同一个样本在不同分布中的状态参数，以此确定是否存在数据异常。图 6 示出根据又一个实施例的比对和判断过程流程图。可以理解，图 6 所示的流程是图 2 中步骤 23 和 24 的子步骤。如图 6 所示，在步骤 61，确定当前获取的多个数据包的数据分布状态作为当前分布；在步骤 62，确定随机抽取的数据包在所述当前分布中的分布状态参数，即第一参数；在步骤 63，确定该随机抽取的数据包在所述历史分布中的分布状态参数，即第二参数；在步骤 64，确定所述第一参数和第二参数之间的差值；然后，在步骤 65，判断所述差值是否超出预定差值阈值，在超出预定差值阈值的情况下，在步骤 66，确定存在数据异常。

[64]以上的流程是随机抽取一个样本，比对该样本在当前分布和历史分布中各自的状态参数，当两者之间的差值超过预定阈值时，则认为当前存在数据异常。下面仍然结合高斯混合模型和聚类算法的例子，描述在不同形式的历史分布的情况下，图 6 流程的具体执行过程。

[65]在一个实施例中，历史分布是采用混合高斯模型对历史数据进行处理得到的历史概率分布。相应地，图 6 的流程可以如下执行。在步骤 61，采用混合高斯模型确定当前获取的多个数据包的混合高斯概率分布作为当前概率分布。在步骤 62，确定随机抽取的数据包在当前概率分布中的出现概率，即第一概率 p_1 ；在步骤 63，确定该随机抽取的数据包在历史概率分布中的出现概率，即第二概率 p_2 。然后，在步骤 64，确定第一概率

p1 和第二概率 p2 之间的概率差值 Δp 。在步骤 65, 判断所述概率差值 Δp 是否超出预定差值阈值, 在超出预定差值阈值的情况下, 确定当前存在数据异常。

[66] 图 7 示出以上比对概率值的示意图。在图 7 中, V1 为示意性示出当前概率分布的曲线, V2 为示意性示出历史概率分布的曲线。可以看到, 随机抽取的数据包 R 在当前概率分布曲线 V1 中的对应概率为 p1, 在历史概率分布曲线 V2 中的对应概率为 p2。如果概率 p1 和 p2 之间的差值 Δp 大于预设阈值, 那么可以认为, 当前概率分布相比于历史概率分布发生了较大变化, 因此当前有可能存在异常。

[67] 可以理解, 混合高斯模型一般基于多维度进行数据统计, 实际的分布曲线图是立体多维的曲线图。图 7 仅仅是一种简化的示意。

10 [68] 以上描述了在高斯分布的情况下, 图 6 所示的比对同一样本的状态参数的执行方式。在历史分布为采用聚类算法获得的聚类分布的情况下, 可以类似地执行图 6 的流程。具体而言, 在步骤 61, 采用相同的聚类算法确定当前获取的多个数据包的聚类分布作为当前聚类分布。在步骤 62, 确定随机抽取的数据包在当前聚类分布中的位置与当前聚类分布对应的聚类的中心之间的距离, 即第一距离 D1; 在步骤 63, 确定该随机抽取的数据包在历史聚类分布中的位置与历史聚类分布对应的聚类的中心之间的距离, 即第二距离 D2。可以理解, 上述随机抽取的数据包可以映射为聚类分布空间中的点, 上述第一距离 D1 和第二距离 D2 为该随机抽取的数据包所映射的点与对应聚类分布的中心之间的距离, 可通过多维空间中多维向量之间的距离计算方式获得。

15 [69] 在步骤 64, 确定所述第一距离 D1 和第二距离 D2 之间的距离差值 ΔD 。在步骤 65, 判断上述距离差值是否超出预定差值阈值, 如果超出该差值阈值, 则确定当前存在数据异常。

[70] 在以上方式中, 比对同一样本在当前聚类分布中远离中心的距离和在历史聚类分布中远离中心的距离, 如果两者出现较大偏差, 则认为当前聚类分布与历史聚类分布存在较大差异, 进而确定当前存在异常。

25 [71] 尽管以上结合混合高斯模型和聚类分布进行了描述, 但是, 历史分布的获得并不局限这两个例子, 还可以采用其他算法, 只要能够分析统计得到历史数据的分布规律即可。此外, 尽管以上结合图 3、图 4 和图 6 的具体流程描述了将当前获取的数据包与历史分布进行比对的方式, 但是可以理解, 比对的方式并不限于这些具体描述的例子。根据历史分布的具体形式, 还可以对其他分布状态参数进行比对。

[72]通过以上实施例描述的方法，即使当前获取的数据包被私密保护而失去业务含义，也可以通过统计历史数据的分布规律，将当前获取的数据包与历史分布进行比对而确定当前是否存在数据异常，从而更有效地发现数据异常，进而进行预警和干预。

[73]根据另一方面的实施例，还提供一种确定数据异常的装置。图 8 示出根据一个实施例的确定装置的示意框图。如图 8 所示，确定装置 800 包括：数据包获取单元 81，配置为获取预定时间段内的多个数据包，所述多个数据包具有相同数据结构；历史获取单元 82，配置为获取具有所述相同数据结构的历史数据的历史分布；比对单元 83，配置为将所述多个数据包与所述历史分布进行比对；确定单元 84，配置为根据比对结果确定是否存在数据异常。

10 [74]在一个实施例中，上述比对单元 83 包括（左侧虚线示出）：代入模块 831，配置为通过将所述多个数据包代入所述历史分布，获取所述多个数据包各自在所述历史分布中的分布状态参数，得到多个分布状态参数；阈值比对模块 832，配置为将所述多个分布状态参数与预定的与分布状态相关的阈值进行比对，确定分布状态参数超出所述阈值的数据包的数目。相应地，确定单元 84 配置为：根据所述分布状态参数超出所述阈值的
15 数据包的数目，确定是否存在数据异常。

[75] 在另一实施例中，上述比对单元 83 包括（右侧虚线示出）：分布确定模块 833，配置为确定所述多个数据包的数据分布状态作为当前分布；分布比对模块 834，配置为将所述当前分布与所述历史分布进行比对。

[76]在一个实施例中，所述分布比对模块 834 配置为：确定所述当前分布的分布中心，
20 以及获取所述历史分布的分布中心；确定所述当前分布的分布中心与所述历史分布的分布中心之间的偏移。相应地，确定单元 84 配置为：响应于所述偏移超出预定偏移阈值，确定存在数据异常。

[77]在一个实施例中，所述分布比对模块 834 配置为：确定随机抽取的数据包在所述当前分布中的分布状态参数，即第一参数；确定该随机抽取的数据包在所述历史分布中的
25 分布状态参数，即第二参数；确定所述第一参数和第二参数之间的差值。相应地，确定单元 84 配置为：响应于所述差值超出预定差值阈值，确定存在数据异常。

[78]在一个实施例中，历史分布是采用混合高斯模型对所述历史数据进行处理得到的历史概率分布。相应地，上述代入模块 831 配置为，通过将所述多个数据包代入所述历史概率分布，获取所述多个数据包各自在所述历史概率分布中的概率值，得到多个概率值；

阈值比对模块 832 配置为, 将所述多个概率值与预定的概率阈值进行比对, 确定概率值小于所述概率阈值的数据包的数目, 即第一数目。确定单元 84 配置为: 响应于所述第一数目超出预定数目阈值, 或者所述第一数目与所述多个数据包的数目之间的比例超出预定比例阈值, 确定存在数据异常。

- 5 [79] 在一个实施例中, 在历史分布为高斯概率分布的情况下, 上述分布确定模块 833 配置为: 采用混合高斯模型对所述多个数据包进行处理得到当前概率分布; 分布比对模块 834 配置为: 将所述当前概率分布与所述历史概率分布进行比对。

- 10 [80] 更具体地, 在一个例子中, 所述分布比对模块 834 配置为: 确定第一峰值位置, 所述第一峰值位置为所述当前概率分布对应的曲线的峰值位置; 获取第二峰值位置, 所述第二峰值位置为所述历史概率分布对应的曲线的峰值位置; 确定所述第一峰值位置和第二峰值位置之间的位置偏移。相应地, 确定单元 84 配置为: 响应于所述位置偏移超出预定偏移阈值, 确定存在数据异常。

- 15 [81] 在另一例子中, 分布比对模块 834 配置为: 确定随机抽取的数据包在所述当前概率分布中的出现概率, 即第一概率; 确定该随机抽取的数据包在所述历史概率分布中的出现概率, 即第二概率; 确定所述第一概率和第二概率之间的概率差值。相应地, 确定单元 84 配置为: 响应于所述概率差值超出预定差值阈值, 确定存在数据异常。

- 20 [82] 在一个实施例中, 历史分布是针对所述历史数据采用聚类算法获得的历史聚类分布。相应地, 上述代入模块 831 配置为: 确定所述多个数据包各自在所述历史聚类分布的分布空间中的位置, 得到多个位置; 确定所述多个位置各自与所述历史聚类分布对应的聚类中心位置之间的距离, 得到多个距离; 阈值比对模块 832 配置为: 将所述多个距离与预设距离阈值进行比对, 确定距离超出所述预设距离阈值的数据包的数目, 即第二数目。对应地, 确定单元 84 配置为: 响应于所述第二数目超出预定数目阈值, 或者所述第二数目与所述多个数据包的数目之间的比例超出预定比例阈值, 确定存在数据异常。

- 25 [83] 在一个实施例中, 在历史分布为历史聚类分布的情况下, 上述分布确定模块 833 配置为: 采用所述聚类算法对所述多个数据包进行处理得到当前聚类分布; 分布比对模块 834 配置为: 将所述当前聚类分布与所述历史聚类分布进行比对。

[84] 更具体地, 在一个例子中, 所述分布比对模块 834 配置为: 确定第一中心位置, 所述第一中心位置为所述当前聚类分布对应的聚类的中心位置; 获取第二中心位置, 所述第二中心位置为所述历史聚类分布对应的聚类的中心位置; 确定所述第一中心位置和第

二中心位置之间的距离。相应地，确定单元 84 配置为：响应于所述距离超出预定距离阈值，确定存在数据异常。

5 [85]在另一例子中，所述分布比对模块 834 配置为：确定随机抽取的数据包在所述当前聚类分布中的位置与所述当前聚类分布对应的聚类的中心之间的距离，即第一距离；确定该随机抽取的数据包在所述历史聚类分布中的位置与所述历史聚类分布对应的聚类的中心之间的距离，即第二距离；确定所述第一距离和第二距离之间的距离差值。相应地，确定单元 84 配置为：响应于所述距离差值超出预定差值阈值，确定存在数据异常。

10 [86]通过以上实施例的装置，即使获取的数据包被私密保护而失去业务含义，也可以通过统计历史数据的分布规律，将当前获取的数据包与历史分布进行比对而确定当前是否存在数据异常，从而更有效地发现数据异常，进而进行预警和干预。

[87]根据另一方面的实施例，还提供一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行结合图 2 至图 6 所描述的方法。

15 [88]根据再一方面的实施例，还提供一种计算设备，包括存储器和处理器，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现结合图 2 至图 6 所述的方法。

[89]本领域技术人员应该可以意识到，在上述一个或多个示例中，本发明所描述的功能可以用硬件、软件、固件或它们的任意组合来实现。当使用软件实现时，可以将这些功能存储在计算机可读介质中或者作为计算机可读介质上的一个或多个指令或代码进行传输。

20 [90]以上所述的具体实施方式，对本发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的具体实施方式而已，并不用于限定本发明的保护范围，凡在本发明的技术方案的基础之上，所做的任何修改、等同替换、改进等，均应包括在本发明的保护范围之内。