(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 111078488 B
(45)授权公告日 2021.11.09

(21)申请号 201811215823.0
(22)申请日 2018.10.18
(65)同一申请的已公布的文献号
申请公布号 CN 111078488 A
(43)申请公布日 2020.04.28
(73)专利权人 杭州海康威视数字技术股份有限公司
地址 310051 浙江省杭州市滨江区阡陌路555号(二期)
(72)发明人 何永健 王辉 李冰杰 徐志威
(74)专利代理机构 北京三高永信知识产权代理有限责任公司 11138
代理人 韩东艳
(51)Int.Cl.
G06F 11/30(2006.01)

(56)对比文件
CN 108075906 A,2018.05.25
CN 107608810 A,2018.01.19
CN 107066365 A,2017.08.18
WO 2008040018 A3,2008.05.22
US 2015199253 A1,2015.07.16

审查员 王丹

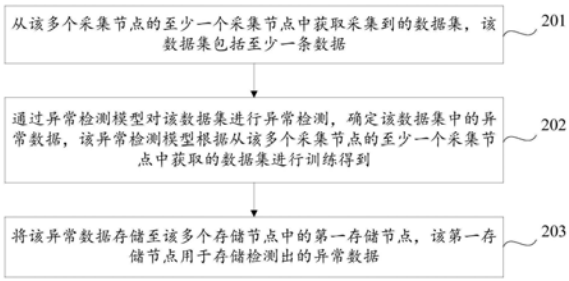
权利要求书5页 说明书17页 附图4页

(54)发明名称

数据采集方法、装置、存储介质及系统

(57)摘要

本发明公开了一种数据采集方法、装置、存储介质及系统，属于大数据技术领域。该方法应用于分布式数据采集系统的指定处理节点中，该方法包括：从多个采集节点的至少一个采集节点中获取采集到的数据集；通过异常检测模型对该数据集进行异常检测，确定该数据集中的异常数据；将该异常数据存储至该多个存储节点中的第一存储节点，该第一存储节点用于存储检测出的异常数据。本发明是根据采集的数据集进行训练得到的异常检测模型，能够反映出区分正常数据和异常数据的规律，学习到正常数据与异常数据之间的区分标准。通过异常检测模型对数据集进行异常检测，能够使得检测结果更加符合真实的异常数据，提高异常检测的准确率。



CN 111078488 B

1.一种数据采集方法,其特征在于,应用于分布式数据采集系统的指定处理节点中,所述分布式数据采集系统包括多个采集节点、多个处理节点和多个存储节点,所述多个处理节点包括一个主处理节点和至少一个从处理节点,所述方法包括:

从所述多个采集节点的至少一个采集节点中获取采集到的数据集,所述数据集包括至少一条数据;通过异常检测模型对所述数据集进行异常检测,确定所述数据集中的异常数据,所述异常检测模型根据从所述多个采集节点的至少一个采集节点中获取的数据集进行训练得到;将所述异常数据存储至所述多个存储节点中的第一存储节点,所述第一存储节点用于存储检测出的异常数据;

当所述指定处理节点为主处理节点时,所述方法还包括:

获取所述主处理节点已训练的异常检测模型,并接收所述至少一个从处理节点已训练的异常检测模型;将所述多个处理节点已训练的异常检测模型进行合成,得到合成后的异常检测模型;将所述合成后的异常检测模型存储至所述多个处理节点共享的存储空间中,供每个处理节点进行异常检测;

当所述指定处理节点为任一从处理节点时,所述方法还包括:

获取所述指定处理节点已训练的异常检测模型,发送给所述主处理节点,所述主处理节点用于将所述多个处理节点已训练的异常检测模型进行合成,得到合成后的异常检测模型,将所述合成后的异常检测模型存储至所述多个处理节点共享的存储空间中,供每个处理节点进行异常检测;

当所述指定处理节点为主处理节点时,所述从所述多个采集节点的至少一个采集节点中获取采集到的数据集之前,所述方法还包括:

监控所述多个采集节点;当监控到所述多个采集节点中的至少一个采集节点采集到数据集时,向至少一个从处理节点发送数据获取指令,所述数据获取指令携带所述至少一个采集节点的所有类型标识或部分类型标识,所述至少一个从处理节点用于根据接收到的数据获取指令中携带的类型标识,从对应的采集节点中获取采集到的数据集;

若所述数据获取指令中携带的类型标识为所述至少一个采集节点的部分类型标识时,所述部分类型标识由所述主处理节点根据类型标识的数量和空闲的处理节点的数量确定。

2.如权利要求1所述的方法,其特征在于,所述多个存储节点包括与所述指定处理节点对应的第二存储节点,所述第二存储节点用于存储所述指定处理节点获取的数据集,所述方法还包括:

从所述第二存储节点中,获取在指定时间段内采集到的数据集,将所述数据集作为第一样本数据集,所述指定时间段是指以开始采集数据的时刻为起始时刻、时间长度为指定时长的时间段;

根据所述第一样本数据集进行训练,得到初始的异常检测模型。

3.如权利要求2所述的方法,其特征在于,所述方法还包括:

从所述第二存储节点中,获取在当前时刻之前的预设时长内采集到的数据集,将所述数据集作为第二样本数据集;

根据所述第二样本数据集继续进行训练,得到更新后的异常检测模型。

4.如权利要求1所述的方法,其特征在于,通过异常检测模型对所述数据集进行异常检测,确定所述数据集中的异常数据,包括:

通过所述异常检测模型对所述数据集进行异常检测，得到所述数据集中每条数据的异常指数，将异常指数处于预设范围内的数据确定为异常数据。

5.如权利要求1所述的方法，其特征在于，所述多个处理节点包括一个主处理节点和至少一个从处理节点，当所述指定处理节点为任一从处理节点时，所述从所述多个采集节点的至少一个采集节点中获取采集到的数据集，包括：

接收所述主处理节点发送的数据获取指令，所述数据获取指令携带至少一个采集节点的类型标识，且所述主处理节点用于当监控到所述至少一个采集节点采集到数据集时发送所述数据获取指令；

根据所述数据获取指令中携带的类型标识，从对应的采集节点中获取采集到的数据集。

6.如权利要求2所述的方法，其特征在于，所述根据所述第一样本数据集进行训练，得到初始的异常检测模型，包括：

根据所述第一样本数据集，建立多个二叉树，将所述多个二叉树进行合成，得到所述初始的异常检测模型；

每个二叉树包括多层节点，第一层中包括一个根节点，每个节点与下一层的两个分支节点连接，每个节点中包括所述第一样本数据集中的一条数据，每个节点的节点值为每个节点中的数据在指定属性上的键值，且每个节点用于将在所述指定属性上的键值小于所述节点值的数据划分至下一层的第一分支节点，将在所述指定属性上的键值不小于所述节点值的数据划分至下一层的第二分支节点。

7.如权利要求6所述的方法，其特征在于，所述根据所述第一样本数据集，建立多个二叉树，包括：

随机从所述第一样本数据集的所有数据属性中选择任一属性，作为指定属性；

随机从所述指定属性的所有键值中选择一个键值作为所述根节点的节点值，将所述节点值对应的数据添加至所述根节点中；

从所述根节点开始，将在所述指定属性上的键值小于当前节点的节点值的数据划分至所述当前节点下一层的第一分支节点，将在所述指定属性上的键值不小于所述当前节点的节点值的数据划分至所述当前节点下一层的第二分支节点，直至划分至的节点中仅包括一条数据或包括在所述指定属性上的键值相同的多条数据时，得到一个二叉树。

8.一种数据采集装置，其特征在于，应用于分布式数据采集系统的指定处理节点中，所述分布式数据采集系统包括多个采集节点、多个处理节点和多个存储节点，所述多个处理节点包括一个主处理节点和至少一个从处理节点，所述装置包括：

第一获取模块，用于从所述多个采集节点的至少一个采集节点中获取采集到的数据集，所述数据集包括至少一条数据；

异常检测模块，用于通过异常检测模型对所述数据集进行异常检测，确定所述数据集中的异常数据，所述异常检测模型根据从所述多个采集节点的至少一个采集节点中获取的数据集进行训练得到；

第一存储模块，用于将所述异常数据存储至所述多个存储节点中的第一存储节点，所述第一存储节点用于存储检测出的异常数据；

当所述指定处理节点为主处理节点时，所述装置还包括：

第四获取模块，用于获取所述主处理节点已训练的异常检测模型，并接收所述至少一个从处理节点已训练的异常检测模型；

合成模块，用于将所述多个处理节点已训练的异常检测模型进行合成，得到合成后的异常检测模型；

第二存储模块，用于将所述合成后的异常检测模型存储至所述多个处理节点共享的存储空间中，供每个处理节点进行异常检测；

当所述指定处理节点为任一从处理节点时，所述装置还包括：

第一发送模块，用于获取所述指定处理节点已训练的异常检测模型，发送给所述主处理节点，所述主处理节点用于将所述多个处理节点已训练的异常检测模型进行合成，得到合成后的异常检测模型，将所述合成后的异常检测模型存储至所述多个处理节点共享的存储空间中，供每个处理节点进行异常检测；

当所述指定处理节点为主处理节点时，所述装置还包括：

监控模块，用于监控所述多个采集节点；

第二发送模块，用于当监控到所述多个采集节点中的至少一个采集节点采集到数据集时，向至少一个从处理节点发送数据获取指令；

所述数据获取指令携带所述至少一个采集节点的所有类型标识或部分类型标识，所述至少一个从处理节点用于根据接收到的数据获取指令中携带的类型标识，从对应的采集节点中获取采集到的数据集；若所述数据获取指令中携带的类型标识为所述至少一个采集节点的部分类型标识时，所述部分类型标识由所述主处理节点根据类型标识的数量和空闲的处理节点的数量确定。

9.如权利要求8所述的装置，其特征在于，所述多个存储节点包括与所述指定处理节点对应的第二存储节点，所述第二存储节点用于存储所述指定处理节点获取的数据集，所述装置还包括：

第二获取模块，用于从所述第二存储节点中，获取在指定时间段内采集到的数据集，将所述数据集作为第一样本数据集，所述指定时间段是指以开始采集数据的时刻为起始时刻、时间长度为指定时长的时间段；

训练模块，用于根据所述第一样本数据集进行训练，得到初始的异常检测模型。

10.如权利要求9所述的装置，其特征在于，所述装置还包括：

第三获取模块，用于从所述第二存储节点中，获取在当前时刻之前的预设时长内采集到的数据集，将所述数据集作为第二样本数据集；

更新模块，用于根据所述第二样本数据集继续进行训练，得到更新后的异常检测模型。

11.如权利要求8所述的装置，其特征在于，所述异常检测模块包括：

异常检测子模块，用于通过所述异常检测模型对所述数据集进行异常检测，得到所述数据集中每条数据的异常指数，将异常指数处于预设范围内的数据确定为异常数据。

12.如权利要求8所述的装置，其特征在于，所述多个处理节点包括一个主处理节点和至少一个从处理节点，当所述指定处理节点为任一从处理节点时，所述第一获取模块包括：

接收子模块，用于接收所述主处理节点发送的数据获取指令，所述数据获取指令携带至少一个采集节点的类型标识，且所述主处理节点用于当监控到所述至少一个采集节点采集到数据集时发送所述数据获取指令；

获取子模块，用于根据所述数据获取指令中携带的类型标识，从对应的采集节点中获取采集到的数据集。

13.如权利要求9所述的装置，其特征在于，所述训练模块包括：

建立子模块，用于根据所述第一样本数据集，建立多个二叉树；

合成子模块，用于将所述多个二叉树进行合成，得到所述初始的异常检测模型；

每个二叉树包括多层节点，第一层中包括一个根节点，每个节点与下一层的两个分支节点连接，每个节点中包括所述第一样本数据集中的一条数据，每个节点的节点值为每个节点中的数据在指定属性上的键值，且每个节点用于将在所述指定属性上的键值小于所述节点值的数据划分至下一层的第一分支节点，将在所述指定属性上的键值不小于所述节点值的数据划分至下一层的第二分支节点。

14.如权利要求13所述的装置，其特征在于，所述建立子模块还用于：

随机从所述第一样本数据集的所有数据属性中选择任一属性，作为指定属性；

随机从所述指定属性的所有键值中选择一个键值作为所述根节点的节点值，将所述节点值对应的数据添加至根节点中；

从所述根节点开始，将在所述指定属性上的键值小于当前节点的节点值的数据划分至所述当前节点下一层的第一分支节点，将在所述指定属性上的键值不小于所述当前节点的节点值的数据划分至所述当前节点下一层的第二分支节点，直至划分至的节点中仅包括一条数据或包括在所述指定属性上的键值相同的多条数据时，得到一个二叉树。

15.一种处理节点，其特征在于，应用于分布式数据采集系统中，所述分布式数据采集系统包括多个采集节点、多个处理节点和多个存储节点，所述处理节点为所述分布式数据采集系统中的任一处理节点；

所述处理节点包括处理器和存储器，所述存储器中存储有至少一条指令，所述至少一条指令由所述处理器加载并执行以实现如权利要求1至7中任一个权利要求所述的数据采集方法。

16.一种计算机可读存储介质，其特征在于，所述存储介质中存储有至少一条指令，所述至少一条指令由处理器加载并执行以实现如权利要求1至7中任一个权利要求所述的数据采集方法。

17.一种分布式数据采集系统，其特征在于，所述分布式数据采集系统包括多个采集节点、多个处理节点和多个存储节点；

所述多个采集节点用于采集数据集，所述数据集包括至少一条数据；

所述多个处理节点中的任一处理节点用于从所述多个采集节点的至少一个采集节点中获取采集到的数据集；

所述任一处理节点还用于通过异常检测模型对所述数据集进行异常检测，确定所述数据集中的异常数据，所述异常检测模型根据从所述多个采集节点的至少一个采集节点中获取的数据集进行训练得到；

所述任一处理节点还用于将所述异常数据存储至所述多个存储节点中的第一存储节点；

所述第一存储节点用于存储检测出的异常数据；

所述多个处理节点包括一个主处理节点和至少一个从处理节点；所述至少一个从处理

节点中的每个从处理节点获取所述每个从处理节点已训练的异常检测模型，发送给所述主处理节点；所述主处理节点用于获取所述主处理节点已训练的异常检测模型，并接收所述至少一个从处理节点已训练的异常检测模型，将所述多个处理节点已训练的异常检测模型进行合成，得到合成后的异常检测模型，将所述合成后的异常检测模型存储至所述多个处理节点共享的存储空间中，供每个处理节点进行异常检测；

当所述任一处理节点为主处理节点时，所述从所述多个采集节点的至少一个采集节点中获取采集到的数据集之前，所述主处理节点用于监控所述多个采集节点，当监控到所述多个采集节点中的至少一个采集节点采集到数据集时，向至少一个从处理节点发送数据获取指令；

所述数据获取指令携带所述至少一个采集节点的所有类型标识或部分类型标识，所述至少一个从处理节点用于根据接收到的数据获取指令中携带的类型标识，从对应的采集节点中获取采集到的数据集；若所述数据获取指令中携带的类型标识为所述至少一个采集节点的部分类型标识时，所述部分类型标识由所述主处理节点根据类型标识的数量和空闲的处理节点的数量确定。

数据采集方法、装置、存储介质及系统

技术领域

[0001] 本发明涉及大数据技术领域，特别涉及一种数据采集方法、装置、存储介质及系统。

背景技术

[0002] 在大数据技术领域中，由于网络的各种原因如网络崩溃、恶意攻击等，可能会产生不符合要求的异常数据，影响数据的后续使用。因此，在数据采集过程中，需要对采集到的数据进行检测，确定其中的异常数据，实现海量数据的有效性检测。

[0003] 相关技术中，通常是预先在分布式数据采集系统中设置固定的预设规则，将预设规则作为正常数据和异常数据的区分标准。那么，分布式数据采集系统每次采集到数据时，按照预设规则对该数据进行区分，将该数据中满足预设规则的数据确定为正常数据，将该数据中不满足预设规则的数据确定为异常数据。

[0004] 在采集海量数据的场景下，随着时间的推移，数据会发生变化，正常数据与异常数据之间的区分标准可能也会发生变化，仍按照上述固定的预设规则进行检测，可能导致无法准确检测出异常数据，从而影响后续工作的正常进行。

发明内容

[0005] 本发明实施例提供了一种数据采集方法、装置、存储介质及系统，可以解决相关技术中采用固定的预设规则进行异常检测，导致无法准确检测出异常数据，影响后续工作正常运行的问题。所述技术方案如下：

[0006] 第一方面，提供了一种数据采集方法，应用于分布式数据采集系统的指定处理节点中，所述分布式数据采集系统包括多个采集节点、多个处理节点和多个存储节点，所述方法包括：

[0007] 从所述多个采集节点的至少一个采集节点中获取采集到的数据集，所述数据集包括至少一条数据；

[0008] 通过异常检测模型对所述数据集进行异常检测，确定所述数据集中的异常数据，所述异常检测模型根据从所述多个采集节点的至少一个采集节点中获取的数据集进行训练得到；

[0009] 将所述异常数据存储至所述多个存储节点中的第一存储节点，所述第一存储节点用于存储检测出的异常数据。

[0010] 可选地，所述多个存储节点包括与所述指定处理节点对应的第二存储节点，所述第二存储节点用于存储所述指定处理节点获取的数据集，所述方法还包括：

[0011] 从所述第二存储节点中，获取在指定时间段内采集到的数据集，将所述数据集作为第一样本数据集，所述指定时间段是指以开始采集数据的时刻为起始时刻、时间长度为指定时长的时间段；

[0012] 根据所述第一样本数据集进行训练，得到初始的异常检测模型。

[0013] 可选地，所述方法还包括：

[0014] 从所述第二存储节点中，获取在当前时刻之前的预设时长内采集到的数据集，将所述数据集作为第二样本数据集；

[0015] 根据所述第二样本数据集继续进行训练，得到更新后的异常检测模型。

[0016] 可选地，所述多个处理节点包括一个主处理节点和至少一个从处理节点，当所述指定处理节点为主处理节点时，所述方法还包括：

[0017] 获取所述主处理节点已训练的异常检测模型，并接收所述至少一个从处理节点已训练的异常检测模型；

[0018] 将所述多个处理节点已训练的异常检测模型进行合成，得到合成后的异常检测模型；

[0019] 将所述合成后的异常检测模型存储至所述多个处理节点共享的存储空间中，供每个处理节点进行异常检测。

[0020] 可选地，所述多个处理节点包括一个主处理节点和至少一个从处理节点，当所述指定处理节点为任一从处理节点时，所述方法还包括：

[0021] 获取所述指定处理节点已训练的异常检测模型，发送给所述主处理节点，所述主处理节点用于将所述多个处理节点已训练的异常检测模型进行合成，得到合成后的异常检测模型，将所述合成后的异常检测模型存储至所述多个处理节点共享的存储空间中，供每个处理节点进行异常检测。

[0022] 可选地，通过异常检测模型对所述数据集进行异常检测，确定所述数据集中的异常数据，包括：

[0023] 通过所述异常检测模型对所述数据集进行异常检测，得到所述数据集中每条数据的异常指数，将异常指数处于预设范围内的数据确定为异常数据。

[0024] 可选地，所述多个处理节点包括一个主处理节点和至少一个从处理节点，当所述指定处理节点为主处理节点时，所述从所述多个采集节点的至少一个采集节点中获取采集到的数据集之前，所述方法还包括：

[0025] 监控所述多个采集节点；

[0026] 当监控到所述多个采集节点中的至少一个采集节点采集到数据集时，向至少一个从处理节点发送数据获取指令，所述数据获取指令携带所述至少一个采集节点的类型标识，所述至少一个从处理节点用于根据接收到的数据获取指令中携带的类型标识，从对应的采集节点中获取采集到的数据集。

[0027] 可选地，所述多个处理节点包括一个主处理节点和至少一个从处理节点，当所述指定处理节点为任一从处理节点时，所述从所述多个采集节点的至少一个采集节点中获取采集到的数据集，包括：

[0028] 接收所述主处理节点发送的数据获取指令，所述数据获取指令携带至少一个采集节点的类型标识，且所述主处理节点用于当监控到所述至少一个采集节点采集到数据集时发送所述数据获取指令；

[0029] 根据所述数据获取指令中携带的类型标识，从对应的采集节点中获取采集到的数据集。

[0030] 可选地，所述根据所述第一样本数据集进行训练，得到初始的异常检测模型，包

括：

[0031] 根据所述第一样本数据集，建立多个二叉树，将所述多个二叉树进行合成，得到所述初始的异常检测模型；

[0032] 每个二叉树包括多层节点，第一层中包括一个根节点，每个节点与下一层的两个分支节点连接，每个节点中包括所述第一样本数据集中的一条数据，每个节点的节点值为每个节点中的数据在指定属性上的键值，且每个节点用于将在所述指定属性上的键值小于所述节点值的数据划分至下一层的第一分支节点，将在所述指定属性上的键值不小于所述节点值的数据划分至下一层的第二分支节点。

[0033] 可选地，所述根据所述第一样本数据集，建立多个二叉树，包括：

[0034] 随机从所述第一样本数据集的所有数据属性中选择任一属性，作为指定属性；

[0035] 随机从所述指定属性的所有键值中选择一个键值作为根节点的节点值，将所述根节点的节点值对应的数据添加至根节点中；

[0036] 从所述根节点开始，将在所述指定属性上的键值小于当前节点的节点值的数据划分至所述当前节点下一层的第一分支节点，将在所述指定属性上的键值不小于所述当前节点的节点值的数据划分至所述当前节点下一层的第二分支节点，直至划分至的节点中仅包括一条数据或包括在所述指定属性上的键值相同的多条数据时，得到一个二叉树。

[0037] 第二方面，提供了一种数据采集装置，应用于分布式数据采集系统的指定处理节点中，所述分布式数据采集系统包括多个采集节点、多个处理节点和多个存储节点，所述装置包括：

[0038] 第一获取模块，用于从所述多个采集节点的至少一个采集节点中获取采集到的数据集，所述数据集包括至少一条数据；

[0039] 异常检测模块，用于通过异常检测模型对所述数据集进行异常检测，确定所述数据集中的异常数据，所述异常检测模型根据从所述多个采集节点的至少一个采集节点中获取的数据集进行训练得到；

[0040] 第一存储模块，用于将所述异常数据存储至所述多个存储节点中的第一存储节点，所述第一存储节点用于存储检测出的异常数据。

[0041] 可选地，所述多个存储节点包括与所述指定处理节点对应的第二存储节点，所述第二存储节点用于存储所述指定处理节点获取的数据集，所述装置还包括：

[0042] 第二获取模块，用于从所述第二存储节点中，获取在指定时间段内采集到的数据集，将所述数据集作为第一样本数据集，所述指定时间段是指以开始采集数据的时刻为起始时刻、时间长度为指定时长的时间段；

[0043] 训练模块，用于根据所述第一样本数据集进行训练，得到初始的异常检测模型。

[0044] 可选地，所述装置还包括：

[0045] 第三获取模块，用于从所述第二存储节点中，获取在当前时刻之前的预设时长内采集到的数据集，将所述数据集作为第二样本数据集；

[0046] 更新模块，用于根据所述第二样本数据集继续进行训练，得到更新后的异常检测模型。

[0047] 可选地，所述多个处理节点包括一个主处理节点和至少一个从处理节点，当所述指定处理节点为主处理节点时，所述装置还包括：

[0048] 第四获取模块，用于获取所述主处理节点已训练的异常检测模型，并接收所述至少一个从处理节点已训练的异常检测模型；

[0049] 合成模块，用于将所述多个处理节点已训练的异常检测模型进行合成，得到合成后的异常检测模型；

[0050] 第二存储模块，用于将所述合成后的异常检测模型存储至所述多个处理节点共享的存储空间中，供每个处理节点进行异常检测。

[0051] 可选地，所述多个处理节点包括一个主处理节点和至少一个从处理节点，当所述指定处理节点为任一从处理节点时，所述装置还包括：

[0052] 第一发送模块，用于获取所述指定处理节点已训练的异常检测模型，发送给所述主处理节点，所述主处理节点用于将所述多个处理节点已训练的异常检测模型进行合成，得到合成后的异常检测模型，将所述合成后的异常检测模型存储至所述多个处理节点共享的存储空间中，供每个处理节点进行异常检测。

[0053] 可选地，所述异常检测模块包括：

[0054] 异常检测子模块，用于通过所述异常检测模型对所述数据集进行异常检测，得到所述数据集中每条数据的异常指数，将异常指数处于预设范围内的数据确定为异常数据。

[0055] 可选地，所述多个处理节点包括一个主处理节点和至少一个从处理节点，当所述指定处理节点为主处理节点时，所述装置还包括：

[0056] 监控模块，用于监控所述多个采集节点；

[0057] 第二发送模块，用于当监控到所述多个采集节点中的至少一个采集节点采集到数据集时，向至少一个从处理节点发送数据获取指令，所述数据获取指令携带所述至少一个采集节点的类型标识，所述至少一个从处理节点用于根据接收到的数据获取指令中携带的类型标识，从对应的采集节点中获取采集到的数据集。

[0058] 可选地，所述多个处理节点包括一个主处理节点和至少一个从处理节点，当所述指定处理节点为任一从处理节点时，所述第一获取模块包括：

[0059] 接收子模块，用于接收所述主处理节点发送的数据获取指令，所述数据获取指令携带至少一个采集节点的类型标识，且所述主处理节点用于当监控到所述至少一个采集节点采集到数据集时发送所述数据获取指令；

[0060] 获取子模块，用于根据所述数据获取指令中携带的类型标识，从对应的采集节点中获取采集到的数据集。

[0061] 可选地，所述训练模块包括：

[0062] 建立子模块，用于根据所述第一样本数据集，建立多个二叉树；

[0063] 合成子模块，用于将所述多个二叉树进行合成，得到所述初始的异常检测模型；

[0064] 每个二叉树包括多层节点，第一层中包括一个根节点，每个节点与下一层的两个分支节点连接，每个节点中包括所述第一样本数据集中的一条数据，每个节点的节点值为每个节点中的数据在指定属性上的键值，且每个节点用于将在所述指定属性上的键值小于所述节点值的数据划分至下一层的第一分支节点，将在所述指定属性上的键值不小于所述节点值的数据划分至下一层的第二分支节点。

[0065] 可选地，所述建立子模块还用于：

[0066] 随机从所述第一样本数据集的所有数据属性中选择任一属性，作为指定属性；

[0067] 随机从所述指定属性的所有键值中选择一个键值作为根节点的节点值，将所述根节点的节点值对应的数据添加至根节点中；

[0068] 从所述根节点开始，将在所述指定属性上的键值小于当前节点的节点值的数据划分至所述当前节点下一层的第一分支节点，将在所述指定属性上的键值不小于所述当前节点的节点值的数据划分至所述当前节点下一层的第二分支节点，直至划分至的节点中仅包括一条数据或包括在所述指定属性上的键值相同的多条数据时，得到一个二叉树。

[0069] 第三方面，提供一种处理节点，应用于分布式数据采集系统中，所述分布式数据采集系统包括多个采集节点、多个处理节点和多个存储节点，所述处理节点为所述分布式采集系统中的任一处理节点；

[0070] 所述处理节点包括处理器和存储器，所述存储器中存储有至少一条指令，所述至少一条指令由所述处理器加载并执行以实现上述第一方面所述的数据采集方法。

[0071] 第四方面，提供一种计算机可读存储介质，所述存储介质中存储有至少一条指令，所述至少一条指令由处理器加载并执行以实现上述第一方面所述的数据采集方法。

[0072] 第五方面，提供一种分布式数据采集系统，所述分布式数据采集系统包括多个采集节点、多个处理节点和多个存储节点；

[0073] 所述多个采集节点用于采集数据集，所述数据集包括至少一条数据；

[0074] 所述多个处理节点中的任一处理节点用于从所述多个采集节点的至少一个采集节点中获取采集到的数据集；

[0075] 所述任一处理节点还用于通过异常检测模型对所述数据集进行异常检测，确定所述数据集中的异常数据，所述异常检测模型根据从所述多个采集节点的至少一个采集节点中获取的数据集进行训练得到；

[0076] 所述任一处理节点还用于将所述异常数据存储至所述多个存储节点中的第一存储节点；

[0077] 所述第一存储节点用于存储检测出的异常数据。

[0078] 本发明实施例中，任一处理节点从该多个采集节点的至少一个采集节点中获取采集到的数据集；然后通过异常检测模型对该数据集进行异常检测，确定该数据集中的异常数据；之后将该异常数据存储至该多个存储节点中的第一存储节点，该第一存储节点用于存储检测出的异常数据。这样，根据采集的数据集进行训练得到的异常检测模型，能够反映出区分正常数据和异常数据的规律，学习到正常数据与异常数据之间的区分标准。那么，任一处理节点使用该异常检测模型对获取到的数据集进行异常检测，确定该数据集中的异常数据并存储，使得检测结果更加符合真实的异常数据，提高了异常检测的准确率，保证了后续工作的正常进行。

附图说明

[0079] 为了更清楚地说明本发明实施例中的技术方案，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他的附图。

[0080] 图1是本发明实施例提供的一种分布式数据采集系统的结构示意图；

[0081] 图2是本发明实施例提供的一种数据采集方法流程图；

[0082] 图3是本发明实施例提供的一种训练异常检测模型的方法流程图；

[0083] 图4是本发明实施例提供的另一种数据采集方法流程图；

[0084] 图5是本发明实施例提供的另一种分布式数据采集系统的结构示意图；

[0085] 图6是本发明实施例提供的一种数据采集装置结构示意图；

[0086] 图7是本发明实施例提供的一种服务器的结构示意图。

具体实施方式

[0087] 为使本发明的目的、技术方案和优点更加清楚，下面将结合附图对本发明实施方式作进一步地详细描述。

[0088] 为了便于理解，在对本发明实施例进行详细地解释说明之前，先对本发明实施例涉及的系统架构进行介绍。

[0089] 图1是本发明实施例提供的一种分布式数据采集系统的结构示意图，参见图1，该分布式数据采集系统中包括多个采集节点101、多个处理节点102和多个存储节点103，至少一个处理节点101与一个处理节点102连接，则每个处理节点102与至少一个采集节点101对应，一个处理节点102与一个存储节点103连接，则该多个处理节点102与该多个存储节点103一一对应。

[0090] 其中，每个采集节点101具备数据采集功能，可以采集数据。每个处理节点102具备异常检测功能，可以对采集到的数据进行异常检测。每个存储节点103具备数据存储功能，可以存储采集到的数据。

[0091] 以多个处理节点102中的任一处理节点为指定处理节点为例，每个采集节点101用于从数据源采集数据集。指定处理节点用于从该多个采集节点101的至少一个采集节点101中获取采集到的数据集，并通过异常检测模型对数据集进行异常检测，得到该数据集中的异常数据，之后将检测出的异常数据存储至第一存储节点。

[0092] 其中，该多个存储节点103中包括该第一存储节点，该第一存储节点用于存储检测出的异常数据。

[0093] 需要说明的是，本发明实施例提供的分布式数据采集系统中包括的采集节点、处理节点和存储节点，可以为服务器，或者也可以为服务器中的功能模块，则不同节点可以部署在同一服务器中，也可以部署在不同服务器上。

[0094] 图2是本发明实施例提供的一种数据采集方法流程图，应用于图1所示的分布式数据采集系统的指定处理节点中，该指定处理节点为分布式数据采集系统中的任一处理节点。参见图2，该方法包括如下步骤：

[0095] 步骤201：从该多个采集节点的至少一个采集节点中获取采集到的数据集，该数据集包括至少一条数据。

[0096] 步骤202：通过异常检测模型对该数据集进行异常检测，确定该数据集中的异常数据，该异常检测模型根据从该多个采集节点的至少一个采集节点中获取的数据集进行训练得到。

[0097] 步骤203：将该异常数据存储至该多个存储节点中的第一存储节点，该第一存储节点用于存储检测出的异常数据。

[0098] 本发明实施例中，任一处理节点从该多个采集节点的至少一个采集节点中获取采集到的数据集；然后通过异常检测模型对该数据集进行异常检测，确定该数据集中的异常数据；之后将该异常数据存储至该多个存储节点中的第一存储节点，该第一存储节点用于存储检测出的异常数据。这样，根据采集的数据集进行训练得到的异常检测模型，能够反映出区分正常数据和异常数据的规律，学习到正常数据与异常数据之间的区分标准。那么，任一处理节点使用该异常检测模型对获取到的数据集进行异常检测，确定该数据集中的异常数据并存储，使得检测结果更加符合真实的异常数据，提高了异常检测的准确率，保证了后续工作的正常进行。

[0099] 图3是本发明实施例提供的一种训练异常检测模型的方法流程图，应用于图1所示实施例的分布式数据采集系统中，该分布式数据采集系统包括多个采集节点、多个处理节点和多个存储节点，该多个处理节点包括一个主处理节点和至少一个从处理节点。参见图3，该方法包括如下步骤：

[0100] 步骤301：主处理节点监控该多个采集节点，当监控到该多个采集节点中的至少一个采集节点采集到数据集时，向至少一个从处理节点发送数据获取指令，该数据获取指令携带该至少一个采集节点的类型标识。

[0101] 其中，采集节点的类型标识用于表示该采集节点的类型，由于在分布式数据采集系统中，类型相同的采集节点之间能够实现数据共享。那么，当主处理节点监控到某一采集节点采集到数据集时，只需要根据采集节点的类型，就可以从该种类型的任一个采集节点中获取采集到的数据集。因此，在该数据获取指令中携带该至少一个采集节点的类型标识即可。该类型标识可以为采集节点的节点类型名称等，如kafka、FTP(File Transfer Protocol，文件传输协议)等。

[0102] 其中，主处理节点可以实时监控每个采集节点，也可以周期性的监控每个采集节点。

[0103] 可选地，该数据获取指令中可以携带该至少一个采集节点的类型标识和采集到的数据集的缓存位置，例如可以携带采集节点的类型标识kafka和数据集的缓存位置topic，以便从处理节点能从kafka采集节点的topic中获取数据。

[0104] 相应地，主处理节点可以监控每个采集节点，在监控到该多个采集节点中的至少一个采集节点采集到数据集时，获取该至少一个采集节点的类型标识和该数据集的缓存位置，并将该至少一个采集节点的类型标识和该数据集的缓存位置携带在数据获取指令中，发送给至少一个从处理节点。

[0105] 当监控到的至少一个采集节点包括多种类型的采集节点时，主处理节点可以在该数据获取指令中携带该至少一个采集节点中的所有类型标识，也可以携带该至少一个采集节点中的部分类型标识。当该数据获取指令中携带的类型标识为部分类型标识时，该部分类型标识由主处理节点根据类型标识的数量和空闲的处理节点的数量确定。

[0106] 可选地，主处理节点确定类型标识的数量和空闲处理节点的数量，计算类型标识的数量和空闲处理节点的数量之间的比值，当该比值小于1时，为至少两个空闲处理节点分配同一个类型标识；当该比值不小于1时，为每个空闲处理节点分配至少一个类型标识，且每个空间处理节点所分配的类型标识不同。

[0107] 其中，空闲处理节点可以为主处理节点和至少一个从处理节点，或者为至少一个

从处理节点，或者为一个主处理节点。

[0108] 例如，若类型标识为2个，当前空闲的处理节点为4个，包括1个主处理节点和3个从处理节点。那么，主处理节点可以为将第一个类型标识指示的采集节点分配给自身和一个从处理节点；将第二个类型标识所指示的采集节点分配给其他两个从处理节点。当然，主处理节点也可以将第一个类型标识指示的采集节点分配给一个从处理节点，将第二个类型标识分配给其他两个从处理节点，而主处理节点本身仅监控采集节点采集数据的情况并进行分配，而不参与获取采集节点的数据集的过程。

[0109] 例如，若类型标识为4个，当前空闲的处理节点为2个从处理节点。主处理节点可以为将两个类型标识分配给一个从处理节点，将另外两个类型标识分配给另一个从处理节点。

[0110] 步骤302：任一处理节点根据该至少一个采集节点的类型标识，从对应的采集节点中获取采集到的数据集并存储至第二存储节点，该任一处理节点为主处理节点和接收到数据获取指令的从处理节点中的任一个。

[0111] 其中，分布式采集系统中，多个存储节点还可以包括多个第二存储节点，每个处理节点对应一个第二存储节点，第二存储节点用于存储对应处理节点获取的数据集。

[0112] 对于主处理节点而言，可以在监控到该至少一个采集节点采集到数据集时，获取该至少一个采集节点的类型标识，根据该类型标识从对应的采集节点中获取采集到的数据集，并将获取到的数据集存储至主处理节点对应的第二存储节点。

[0113] 对于每个从处理节点而言，可以在接收到一个数据获取指令时，就根据该数据获取指令中携带的类型标识，确定要获取其数据集的采集节点，从对应的采集节点中获取采集到的数据集，并将获取到的数据集存储至从处理节点对应的第二存储节点。当然，也可以在接收到的数据获取指令的个数达到预设数量时，对多个数据获取指令进行统一处理。或者，也可以在接收到第一个数据获取指令时，开始计时，当到达一定时间间隔时，根据该时间间隔内接收到的数据获取指令中携带的类型标识，从对应的采集节点中获取采集到的数据集，并重新开始计时。

[0114] 对于采集节点而言，一个采集节点每次只允许一个处理节点获取采集的数据集，多个处理节点不能同时获取同一个采集节点的数据集。那么当一个处理节点在获取某一采集节点采集到的数据集时，其他处理节点则无法获取该采集节点采集到的数据集，只能再去获取其他采集节点采集到的数据集或者不再获取数据集。通过这种方式，使得每个处理节点获取不同采集节点采集到的数据集，这样可以避免多个处理节点获取相同的数据集。

[0115] 需要说明的是，对于每个从处理节点而言，无论该数据获取指令中携带的是该至少一个采集节点中所有采集节点的类型标识，还是携带该至少一个采集节点中部分采集节点的类型标识，从处理节点只需按照该数据获取指令中携带的类型标识从对应的采集节点中获取数据集。

[0116] 而从处理节点根据该数据获取指令中携带的某一个类型标识，从对应的采集节点获取采集到的数据集时，可能存在其他处理节点正在获取该采集节点采集到的数据集，此时，该从处理节点从与该类型标识对应的其他采集节点获取采集到的数据集。对于该类型标识对应的所有采集节点中的每个采集节点，若均存在其他处理节点正在获取采集到的数据集，此时，从处理节点停止根据该类型标识从对应的采集节点获取采集到的数据集。之

后,从处理节点再根据该类型标识的下一个类型标识从对应的采集节点获取采集到的数据集;若该数据获取指令中仅携带了这一个类型标识,则从处理节点停止数据集的获取。

[0117] 步骤303:任一处理节点从对应的第二存储节点中获取在指定时间段内采集到的数据集,将该数据集中的数据作为第一样本数据集,该指定时间段是指以开始采集数据的时刻为起始时刻、时间长度为指定时长的时间段。

[0118] 其中,该指定时长可以设置为一天、两天、12小时等,也可以设置为其他时长。例如,该处理节点可以为Spark Streaming组件,该采集节点可以为Kafka,Kafka采集到的数据集缓存在对应的topic主题中,任一Spark Streaming组件根据采集节点的类型标识Kafka和数据集的存储位置topic1,从Kafka的topic1中获取数据集,然后将该第一样本数据集封装到DStream(数据流)中,之后通过遍历DStream中的RDD(Resilient Distributed Datasets,弹性分布式数据集),来获取每条数据进行后续的模型训练。

[0119] 例如,某一处理节点获取到的第一样本数据集可以如下表1所示。

[0120] 表1

[0121]

	Number_count	Id_count	p_count	Map_agent_count	Name_len_count	prefix_count	price
0	0	0	0	0	0	0	0
1	0.744680851	0.744681	0.065217	0.0315	0.571681286	0.21965486	56.04167
2	0.021270596	0.021277	0.021739	0.0315	0	0	79.89256

[0122]

3	0.021270596	0.021277	0.021739	0	0.028571429	0	4.165
4	0.021270596	0.021277	0	0	0.028571429	0	67.4862
5	0.042625844	0.06383	0.021739	0	0	0.0216325	67.1586
6	0.021270596	0.021277	0.043478	0	0.084625962	0	56.41654
7	0.021270596	0.042553	0.021739	0.0315	0.028571429	0.04929858	66.469
8	0.064825843	0.021277	0	0	0.028571429	0.0216325	157.655

[0123] 需要说明的是,上述表1仅仅是本发明实施例提供的一种示例性的第一样本数据集,该第一样本数据集也可以为其他,对此本发明实施例不予限定。

[0124] 步骤304:任一处理节点根据第一样本数据集进行训练,得到初始的异常检测模型。

[0125] 其中,该任一处理节点可以为一个主处理节点和至少一个从处理节点。因此,对于一个主处理节点和至少一个从处理节点中的每个处理节点而言,都会获取到一个第一样本数据集进行训练,得到一个初始的异常检测模型,那么,最后会得到多个初始的异常检测模型。并且不同处理节点获取的第一样本数据集为不同数据集的数据,使得获取的数据更全面。

[0126] 该异常检测模型是直接根据获取的数据集进行训练得到,更符合数据集本身的区分标准,且该异常检测模型不需要预先设置区分标准,适用于海量数据集的异常检测,准确率高。其中,该异常检测模型可以基于Isolation-Forest(孤立森林)算法实现。

[0127] 任一处理节点在获取到第一样本数据集时,可以先根据第一样本数据集,建立多

个二叉树，然后将该多个二叉树进行合成，得到初始的异常检测模型。

[0128] 其中，每个二叉树包括多层节点，第一层中包括一个根节点，每个节点与下一层的两个分支节点连接，每个节点中包括第一样本数据集中的一条数据，每个节点的节点值为每个节点中的数据在指定属性上的键值，且每个节点用于将在指定属性上的键值小于该节点值的数据划分至下一层的第一分支节点，将在指定属性上的键值不小于该节点值的数据划分至下一层的第二分支节点。

[0129] 可选地，任一处理节点根据第一样本数据集，建立多个二叉树时，可以随机从第一样本数据集的所有数据属性中选择任一属性，作为指定属性；随机从指定属性的所有键值中选择一个键值作为根节点的节点值，将根节点的节点值对应的数据添加至根节点中；从根节点开始，将在指定属性上的键值小于当前节点的节点值的数据划分至当前节点下一层的第一分支节点，将在指定属性上的键值不小于当前节点的节点值的数据划分至当前节点下一层的第二分支节点，直至划分至的节点中仅包括一条数据或包括在指定属性上的键值相同的多条数据时，得到一个二叉树。

[0130] 建立一个二叉树的代码可以如下所示：

[0131]
```
def growTree(X-data, e-currentHeight, l-heightLimit)
if e ⩾ l or |X| ⩽ 1 then
    return exNode {Size ← |X|}
else
    X_l ← filter(X, q<p)
    X_r ← filter(X, q ⩾ p)
    Return inNode{Left ← iTree(X_l,e+1,l)
                  Right ← iTree(X_r,e+1,l)
                  SplitAtt ← q,
                  SplitValue ← p}
```

[0132] 其中，Att为指定属性，Value为节点值，X为随机选取的指定属性的所有键值，e为当前高度，l为指定高度，该指定高度为预先设置。

[0133] 需要说明的是，若随机选取的某个节点的节点值高于指定高度，或者不存在指定属性上的键值小于该节点值的数据，或者不存在指定属性上的键值不小于该节点值的数据，则表示选取的该节点值不合适，返回重新选取该节点的节点值。

[0134] 步骤305：每个从处理节点获取已训练的异常检测模型，发送给主处理节点。

[0135] 步骤306：主处理节点获取当前已训练的异常检测模型，并接收至少一个从处理节点已训练的异常检测模型。

[0136] 步骤307：主处理节点将多个处理节点已训练的异常检测模型进行合成，得到合成后的异常检测模型。

[0137] 主处理节点将自身当前已训练的异常检测模型和接收到的至少一个从处理节点已训练的异常检测模型进行合成处理，最终得到一个合成后的异常检测模型。

[0138] 也即是说，主处理节点将多个处理节点根据不同数据集训练得到的异常检测模型进行合成，能够综合考虑该至少一个采集节点在指定时间段内采集到的所有数据集，使得最终得到的合成后的异常检测模型更能体现出数据集本身的规律，保证了异常检测模型的准确率。

[0139] 步骤308：主处理节点将合成后的异常检测模型存储至该多个处理节点共享的存储空间中，供每个处理节点进行异常检测。

[0140] 其中，该多个处理节点共享的存储空间，可以位于主处理节点，也可以位于其他处理节点，或者位于一个独立可供该多个处理节点共同访问的存储服务器上。

[0141] 考虑到在采集海量数据的场景下，随着时间的推移，正常数据与异常数据之间的区分标准可能会发生变化，因此，本发明实施例可以通过下述方式对异常检测模型进行更新，以使更新后的异常检测模型更加符合当前区分正常数据和异常数据的规律，从而能够提高异常检测的准确率。

[0142] 在按照上述步骤301-308建立异常检测模型之后，上述方法还包括：任一处理节点从第二存储节点中，获取在当前时刻之前的预设时长内采集到的数据集，将该数据集作为第二样本数据集；根据该第二样本数据集继续进行训练，得到新的异常检测模型作为更新后的异常检测模型。

[0143] 之后按照上述步骤305-308的方法由每个从处理节点获取当前已训练的异常检测模型，发送给主处理节点。主处理节点获取当前已训练的异常检测模型，并接收至少一个从处理节点已训练的异常检测模型，然后将多个处理节点已训练的异常检测模型进行合成，得到合成后的异常检测模型。之后，将当前该多个处理节点共享的存储空间中存储的异常检测模型替换为该合成后的异常检测模型，以供每个处理节点进行异常检测。

[0144] 需要说明的是，为了保证异常检测模型的准确性，可以周期性的对异常检测模型进行更新，例如，可以在处理节点中设置在每天零点获取零点之前预设时长内采集到的数据集，对异常检测模型进行更新，或者设置其他时间定时对异常检测模型进行更新，其中，该更新周期可以为与上述预设时长相同，也可以不同。这样，使用新采集的数据集进行训练得到的异常检测模型，替换该多个处理节点共享的存储空间中当前存储的异常检测模型，实现对异常检测模型的更新，从而保证在数据不断变化时，能够根据数据的变化情况，更新多个处理节点共享的存储空间中当前存储的异常检测模型，从而提高异常检测的准确度。

[0145] 其中，每个处理节点对应该多个存储节点中的一个第二存储节点，当任一处理节点在从该至少一个采集节点获取采集到的数据集时，可以将获取到的数据集存储至该处理节点对应的第二存储节点，以将未处理的数据集进行备份，用于后续对异常检测模型进行更新，也可以用于进行其他处理。

[0146] 另外，图3实施例提供的方法应用于图1所示实施例的分布式数据采集系统中，该分布式数据采集系统中的指定处理节点可以为主处理节点，此时，该指定处理节点用于执行上述步骤301-308中主处理节点执行的操作，通过与从处理节点之间的交互，执行上述图3实施例所示的方法。该分布式数据采集系统中的指定处理节点也可以为任一从处理节点，此时，该指定处理节点用于执行上述步骤301-308中从处理节点执行的操作，通过与主处理节点之间的交互，执行上述图3实施例所示的方法。

[0147] 综上所述，本发明实施例中，通过主处理节点对多个采集节点进行监控，以使多个

处理节点的任一处理节点从多个采集节点中至少一个采集节点中获取采集到的数据集，直接使用获取到的数据集进行训练得到的异常检测模型，能够反映出区分正常数据和异常数据的规律，学习到正常数据与异常数据之间的区分标准。并且，还可以在后续数据不断变化时，能够根据数据的变化情况，更新多个处理节点共享的存储空间中当前存储的异常检测模型，从而提高异常检测的准确度。

[0148] 图4是本发明实施例提供的一种数据采集方法的流程图，应用于图1所示实施例的分布式数据采集系统中，该分布式数据采集系统包括多个采集节点、多个处理节点和多个存储节点，该多个处理节点包括一个主处理节点和至少一个从处理节点。参见图4，该方法包括如下步骤：

[0149] 步骤401：主处理节点监控该多个采集节点，当监控到该多个采集节点中的至少一个采集节点采集到数据集时，向至少一个从处理节点发送数据获取指令，该数据获取指令携带该至少一个采集节点的类型标识。

[0150] 步骤402：任一处理节点根据该至少一个采集节点的类型标识，从对应的采集节点中获取采集到的数据集，该处理节点为主处理节点和接收到数据获取指令的从处理节点中的任一个，该数据集包括至少一条数据。

[0151] 步骤403：任一处理节点从该多个处理节点共享的存储空间中获取异常检测模型，并通过异常检测模型对该数据集进行异常检测，确定该数据集中的异常数据。

[0152] 任一处理节点在获取到数据集时，可以从该多个处理节点共享的存储空间中获取异常检测模型，该异常检测模型为上述图3实施例训练后得到并存储的。

[0153] 任一处理节点在获取到异常检测模型后，通过异常检测模型对该数据集进行异常检测，得到该数据集中每条数据的异常指数，将异常指数处于预设范围内的数据确定为异常数据。

[0154] 需要说明的是，任一处理节点在获取到该数据集，使用异常检测模型对该数据集进行异常检测，该异常检测模型中包括多个二叉树，而每个二叉树按照一个属性，通过每个节点的节点值对该属性的键值进行划分，因此，在对该数据集中的每条数据进行异常检测时，将数据的指定属性的键值输入到异常检测模型中，通过每个二叉树对该键值进行层层划分，直至将该键值划分至二叉树的最后一层分支节点，记录这个键值的路径长度，即树高度。根据如下公式(1)-(3)计算出数据该键值的异常指数，按照同样的方式计算出该条数据的所有键值的异常指数，根据该条数据中的所有键值的异常指数，确定该条数据的异常指数，若该条数据的异常指数处于预设范围内，则确定该条数据为异常数据。

[0155] $h(x)=e+c(n)$ 公式(1)

[0156] 其中，e表示数据x从二叉树的根节点到最后一层分支节点过程中经过的节点数目，c(n)为修正项；

[0157] $c(n)=2H(n-1)-(2(n-1)/n)$； 公式(2)

[0158] 其中，$H(k)=\ln(k)+\xi$，ξ为欧拉常数，取值为0.5772156649；

[0159] $$S(x,n)=2^{-\frac{h(x)}{c(n)}}$$ 公式(3)

[0160] 其中，S(x,n)用于表示当前数据的异常指数；

[0161] $$S(x,n)=\begin{cases}S(x,n)\rightarrow 1\\ S(x,n)\rightarrow 0\end{cases}\quad S(x,n)\in[0,1]$$

[0162] S(x,n)→1表示数据X异常的可能性越大,S(x,n)→0表示数据X异常的可能性越小。当该数据集中大多数数据的异常指数S(x,n)都接近于0.5,说明整个数据集没有明显的异常情况。

[0163] 一种可能实现的方式,上述确定某条数据的异常指数时,可以通过计算该条数据的所有键值的异常指数的平均值,将该平均值作为该条数据的异常指数,然后确定该条数据的异常指数是否位于预设范围,若是,则确定该条数据异常。

[0164] 其中,该预设范围可以预先设置,如(0.6,1),当然也可以为其他,如(0.8,1)。

[0165] 在本发明实施例中,任一处理节点采用上述方式,通过异常检测模型对获取的数据集中的每一条数据进行异常检测,快速确定该数据集中的所有数据的异常指数,进而确定该数据集中的异常数据,实现对数据集的异常检测。

[0166] 步骤404:任一处理节点将该异常数据存储至该多个存储节点中的第一存储节点,该第一存储节点用于存储检测出的异常数据。

[0167] 每个处理节点对应该多个存储节点中的一个第一存储节点,当任一处理节点得到异常数据时,可以将该异常数据存储至自身对应的第一存储节点中,该第一存储节点可以将检测出的异常数据存储,以供用户后续检索异常数据,以及分析异常原因。

[0168] 例如,该第一存储节点可以为Elasticsearch组件,那么用户可以通过Elasticsearch组件对异常数据进行检索,并对检索结果进行分析。Elasticsearch组件可以在存储异常数据时,为异常数据设置用于存储和检索的索引名。例如,异常数据的索引名可以为anormal,当然也可以为其他。用户可以根据索引名对异常数据进行检索,得到的部分异常数据的检索结果可以如下所示:

{

“_index”: “anormal”,

“_type”: “item”,

“_id”: “AWUTmL_XiommX4iKIZq7”,

“_score”: 1,

“_score”: {“_data”: [0.744680851, 0.744680851, 0.0246956, 0, 1,

[0169] 0.71658592, 115], “anormalscore” :“0.746989258992569”}

}

{

“_index”: “anormal”,

“_type”: “item”,

“_id”: “AWUTfrijefhihji235_ht”,

“_score”: 1,

"_score": {"_data": [1， 1，0,0.29852985985，0，0.418926478，0.0246956，

[0170] 1，0]，"anormalscore" :"0.7879261458"}

}

[0171] 其中,anormal表示异常数据的索引名,type为索引类型,_id为当前条数据的唯一标识符,_source中的内容为获取到的数据集中一条数据data和该数据中的异常指数anormalScore。

[0172] 一种可能的实现方式中,每个处理节点可以将数据集中的正常数据与异常数据分开存储在第一存储节点。

[0173] 相应地,可以为正常数据和异常数据设置不同的索引名,以用于对正常数据和异常数据进行检索。如设置异常数据的索引名为anormal,正常数据的索引名可以为normal,当然也可以设置为其他。

[0174] 需要说明的是,图4实施例提供的方法应用于图1所示实施例的分布式数据采集系统中,该分布式数据采集系统中的指定处理节点可以为主处理节点,此时,该指定处理节点用于执行上述步骤401-405中主处理节点执行的操作,通过与从处理节点之间的交互,执行上述图4实施例所示的方法。该分布式数据采集系统中的指定处理节点也可以为从处理节点,此时,该指定处理节点用于执行上述步骤401-405中从处理节点执行的操作,通过与主处理节点之间的交互,执行上述图4实施例所示的方法。

[0175] 一种可能实现的方式中,本发明实施例的分布式数据采集系统可以包括多个kafka模块、多个Spark Streaming模块、多个Elasticsearch模块和多个Hbase模块,该多个Spark Streaming模块包括一个主Spark Streaming模块和至少一个从Spark Streaming模块。上述图3和图4实施例中采集节点所执行的操作可以由kafka模块执行。上述图3和图4实施例中处理节点所执行的操作可以由SparkStreaming模块来执行。上述图3和图4实施例中第一存储节点所执行的操作可以由Elasticsearch模块来执行,且该Elasticsearch模块还可以提供分析以及搜索功能,从而便于用户对异常数据进行搜索和分析,以评估数据异常原因,便于进行后续的工作。上述图3和图4实施例中第二存储节点所执行的操作可以由Hbase模块来执行。

[0176] 接下来,如图5所示,以该分布式数据采集系统包括一个kafka模块、一个Spark Streaming模块、一个Elasticsearch模块和一个Hbase模块为例进行说明。

[0177] Spark Streaming模块监控kafka模块,当Spark Streaming模块监控到kafka模块采集到数据集时,Spark Streaming模块从kafka模块中获取采集到的数据集,并将获取到的数据集存储至Hbase模块。然后Spark Streaming模块使用预先根据从kafka模块中获取的数据集进行训练得到的异常检测模型,对当前获取到的数据集进行异常检测,确定该数据集中的异常数据。之后Spark Streaming模块将该异常数据存储至Elasticsearch模块。

[0178] 综上所述,本发明实施例中,主处理节点监控多个采集节点,使得该多个处理节点能够在该多个采集节点的至少一个采集节点采集到数据集时,获取采集的数据集,然后每个处理节点从共享的存储空间中获取异常检测模型,并使用异常检测模型对该数据集进行异常检测,确定该数据集中的异常数据;之后将该异常数据存储至该多个存储节点中的第一存储节点,以供后续使用。在本发明的分布式数据采集系统中,异常检测模型能够反映出区分正常数据和异常数据的规律,多个处理节点使用共享的存储空间中存储的异常检测模

型，能够并行对海量数据集进行异常检测，保证了异常检测的准确度的同时，还提高了异常检测的速度。并且，本发明实施例提供的数据采集方法能够在后续数据不断变化时，能够根据数据的变化情况，更新多个处理节点共享的存储空间中当前存储的异常检测模型，通过更新后的异常检测模型进行异常检测，使得异常检测结果更加准确。

[0179] 图6是本发明实施例提供的一种数据采集装置的结构示意图。参见图6，该装置应用于分布式数据采集系统的指定处理节点中，该分布式数据采集系统包括多个采集节点、多个处理节点和多个存储节点，该装置包括第一获取模块601、异常检测模块602和第一存储模块603。

[0180] 第一获取模块601，用于从该多个采集节点的至少一个采集节点中获取采集到的数据集，该数据集包括至少一条数据；

[0181] 异常检测模块602，用于通过异常检测模型对该数据集进行异常检测，确定该数据集中的异常数据，该异常检测模型根据从该多个采集节点的至少一个采集节点中获取的数据集进行训练得到；

[0182] 第一存储模块603，用于将该异常数据存储至该多个存储节点中的第一存储节点，第一存储节点用于存储检测出的异常数据。

[0183] 可选地，该多个存储节点包括与指定处理节点对应的第二存储节点，第二存储节点用于存储指定处理节点获取的数据，该装置还包括：

[0184] 第二获取模块，用于从第二存储节点中，获取在指定时间段内采集到的数据集，将该数据集作为第一样本数据集，该指定时间段是指以开始采集数据的时刻为起始时刻、时间长度为指定时长的时间段；

[0185] 训练模块，用于根据第一样本数据集进行训练，得到初始的异常检测模型。

[0186] 可选地，该装置还包括：

[0187] 第三获取模块，用于从第二存储节点中，获取在当前时刻之前的预设时长内采集到的数据集，将该数据集作为第二样本数据集；

[0188] 更新模块，用于根据第二样本数据集继续进行训练，得到更新后的异常检测模型。

[0189] 可选地，该多个处理节点包括一个主处理节点和至少一个从处理节点，当该指定处理节点为主处理节点时，该装置还包括：

[0190] 第四获取模块，用于获取主处理节点已训练的异常检测模型，并接收该至少一个从处理节点已训练的异常检测模型；

[0191] 合成模块，用于将该多个处理节点已训练的异常检测模型进行合成，得到合成后的异常检测模型；

[0192] 第二存储模块，用于将合成后的异常检测模型存储至该多个处理节点共享的存储空间中，供每个处理节点进行异常检测。

[0193] 可选地，该多个处理节点包括一个主处理节点和至少一个从处理节点，当该指定处理节点为任一从处理节点时，该装置还包括：

[0194] 第一发送模块，用于获取该指定处理节点已训练的异常检测模型，发送给主处理节点，主处理节点用于将该多个处理节点已训练的异常检测模型进行合成，得到合成后的异常检测模型，将合成后的异常检测模型存储至该多个处理节点共享的存储空间中，供每个处理节点进行异常检测。

[0195] 可选地，异常检测模块602包括：

[0196] 异常检测子模块，用于通过异常检测模型对该数据集进行异常检测，得到该数据集中每条数据的异常指数，将异常指数处于预设范围内的数据确定为异常数据。

[0197] 可选地，该多个处理节点包括一个主处理节点和至少一个从处理节点，当指定处理节点为主处理节点时，该装置还包括：

[0198] 监控模块，用于监控该多个采集节点；

[0199] 第二发送模块，用于当监控到该多个采集节点中的至少一个采集节点采集到数据集时，向至少一个从处理节点发送数据获取指令，该数据获取指令携带该至少一个采集节点的类型标识，该至少一个从处理节点用于根据接收到的数据获取指令中携带的类型标识，从对应的采集节点中获取采集到的数据集。

[0200] 可选地，该多个处理节点包括一个主处理节点和至少一个从处理节点，当指定处理节点为任一从处理节点时，第一获取模块包括：

[0201] 接收子模块，用于接收主处理节点发送的数据获取指令，该数据获取指令携带至少一个采集节点的类型标识，且主处理节点用于当监控到该至少一个采集节点采集到数据集时发送该数据获取指令；

[0202] 获取子模块，用于根据该数据获取指令中携带的类型标识，从对应的采集节点中获取采集到的数据集。

[0203] 可选地，训练模块包括：

[0204] 建立子模块，用于根据第一样本数据集，建立多个二叉树；

[0205] 合成子模块，用于将该多个二叉树进行合成，得到初始的异常检测模型；

[0206] 每个二叉树包括多层节点，第一层中包括一个根节点，每个节点与下一层的两个分支节点连接，每个节点中包括第一样本数据集中的一条数据，每个节点的节点值为每个节点中的数据在指定属性上的键值，且每个节点用于将在指定属性上的键值小于该节点值的数据划分至下一层的第一分支节点，将在指定属性上的键值不小于该节点值的数据划分至下一层的第二分支节点。

[0207] 可选地，建立子模块还用于：

[0208] 随机从第一样本数据集的所有数据属性中选择任一属性，作为指定属性；

[0209] 随机从指定属性的所有键值中选择一个键值作为根节点的节点值，将根节点的节点值对应的数据添加至根节点中；

[0210] 从根节点开始，将在指定属性上的键值小于当前节点的节点值的数据划分至当前节点下一层的第一分支节点，将在指定属性上的键值不小于当前节点的节点值的数据划分至当前节点下一层的第二分支节点，直至划分至的节点中仅包括一条数据或包括在该指定属性上的键值相同的多条数据时，得到一个二叉树。

[0211] 本发明实施例中，任一处理节点从该多个采集节点的至少一个采集节点中获取采集到的数据集；然后通过异常检测模型对该数据集进行异常检测，确定该数据集中的异常数据；之后将该异常数据存储至该多个存储节点中的第一存储节点，该第一存储节点用于存储检测出的异常数据。这样，根据采集的数据集进行训练得到的异常检测模型，能够反映出区分正常数据和异常数据的规律，学习到正常数据与异常数据之间的区分标准。那么，任一处理节点使用该异常检测模型对获取到的数据集进行异常检测，确定该数据集中的异常

数据并存储，使得检测结果更加符合真实的异常数据，提高了异常检测的准确率，保证了后续工作的正常进行。

[0212] 需要说明的是：上述实施例提供的数据采集装置在采集数据时，仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将装置的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的数据采集装置与数据采集方法实施例属于同一构思，其具体实现过程详见方法实施例，这里不再赘述。

[0213] 图7是本发明实施例提供的一种服务器的结构示意图，该服务器700可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上处理器(central processing units，CPU)701和一个或一个以上的存储器702，其中，所述存储器702中存储有至少一条指令，所述至少一条指令由该处理器701加载并执行。当然，该服务器700还可以具有有线或无线网络接口、键盘以及输入输出接口等部件，以便进行输入输出，该服务器700还可以包括其他用于实现设备功能的部件，在此不做赘述。

[0214] 该服务器700用于执行上述数据获取方法中控制设备或节点设备所执行的操作。

[0215] 在示例性实施例中，还提供了一种计算机可读存储介质，例如包括指令的存储器，上述指令可由上述终端或服务器中的处理器执行以完成上述实施例中的数据采集方法。例如，所述计算机可读存储介质可以是ROM、随机存取存储器(RAM)、CD-ROM、磁带、软盘和光数据存储设备等。

[0216] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成，也可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，上述提到的存储介质可以是只读存储器，磁盘或光盘等。

[0217] 以上所述仅为本发明的较佳实施例，并不用以限制本发明，凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

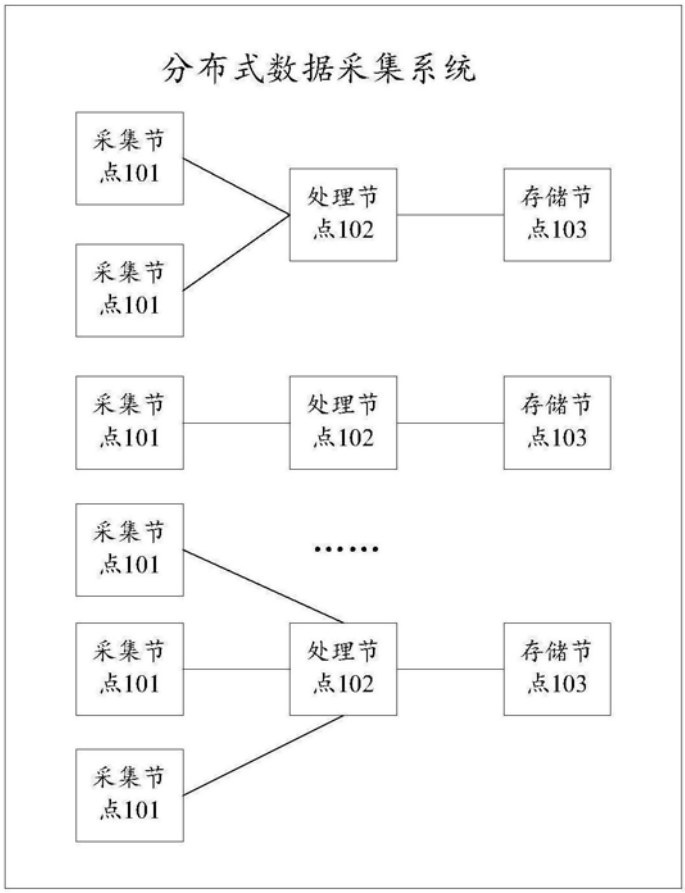

分布式数据采集系统
采集节点101
采集节点101
处理节点102
存储节点103
采集节点101
处理节点102
存储节点103
采集节点101
……
采集节点101
处理节点102
存储节点103
采集节点101

图1

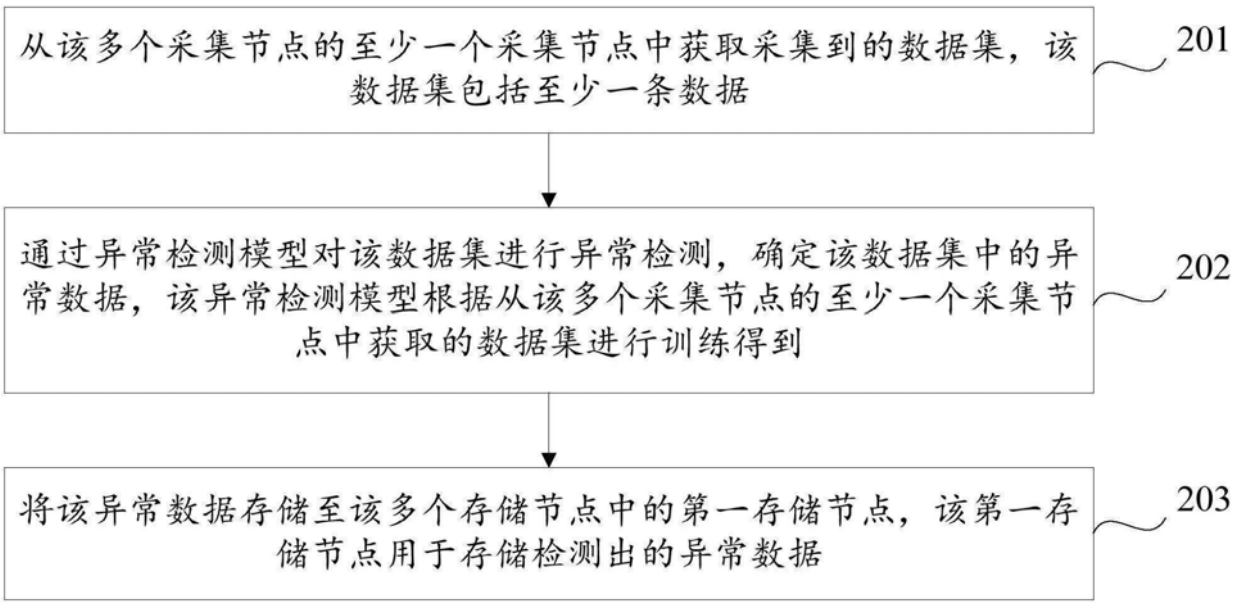

从该多个采集节点的至少一个采集节点中获取采集到的数据集，该数据集包括至少一条数据
201
通过异常检测模型对该数据集进行异常检测，确定该数据集中的异常数据，该异常检测模型根据从该多个采集节点的至少一个采集节点中获取的数据集进行训练得到
202
将该异常数据存储至该多个存储节点中的第一存储节点，该第一存储节点用于存储检测出的异常数据
203

图2

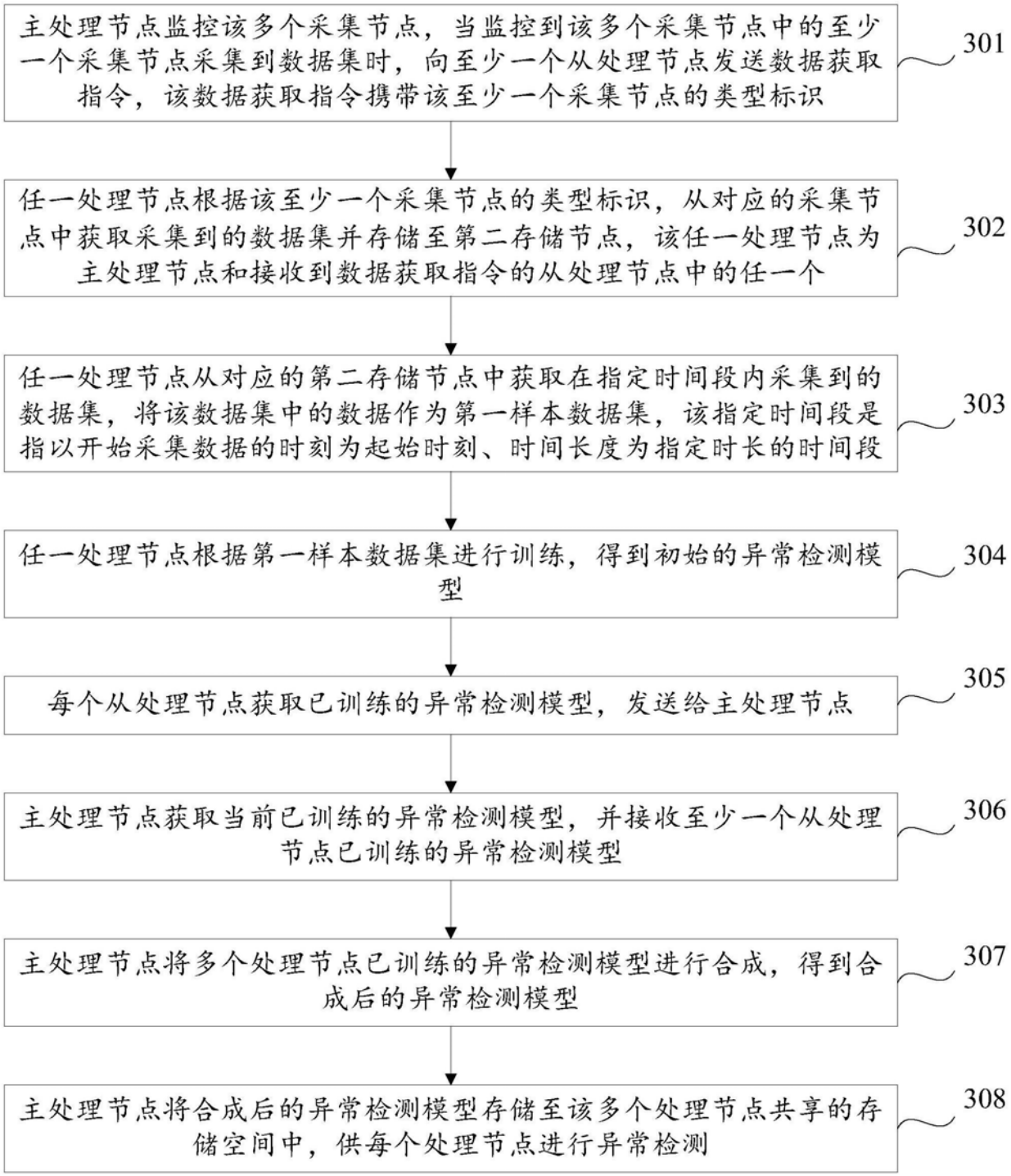

主处理节点监控该多个采集节点，当监控到该多个采集节点中的至少一个采集节点采集到数据集时，向至少一个从处理节点发送数据获取指令，该数据获取指令携带该至少一个采集节点的类型标识
301
任一处理节点根据该至少一个采集节点的类型标识，从对应的采集节点中获取采集到的数据集并存储至第二存储节点，该任一处理节点为主处理节点和接收到数据获取指令的从处理节点中的任一个
302
任一处理节点从对应的第二存储节点中获取在指定时间段内采集到的数据集，将该数据集中的数据作为第一样本数据集，该指定时间段是指以开始采集数据的时刻为起始时刻、时间长度为指定时长的时间段
303
任一处理节点根据第一样本数据集进行训练，得到初始的异常检测模型
304
每个从处理节点获取已训练的异常检测模型，发送给主处理节点
305
主处理节点获取当前已训练的异常检测模型，并接收至少一个从处理节点已训练的异常检测模型
306
主处理节点将多个处理节点已训练的异常检测模型进行合成，得到合成后的异常检测模型
307
主处理节点将合成后的异常检测模型存储至该多个处理节点共享的存储空间中，供每个处理节点进行异常检测
308

图3

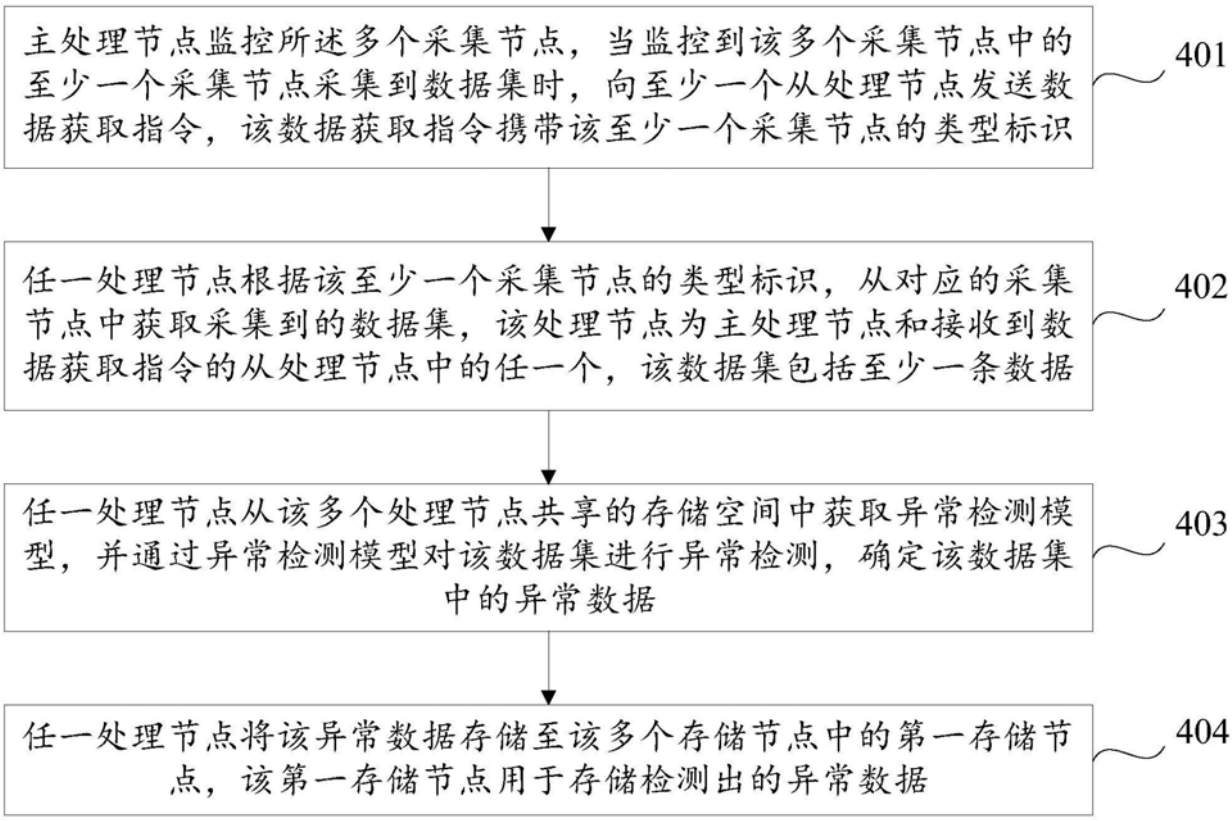

主处理节点监控所述多个采集节点，当监控到该多个采集节点中的至少一个采集节点采集到数据集时，向至少一个从处理节点发送数据获取指令，该数据获取指令携带该至少一个采集节点的类型标识
401
任一处理节点根据该至少一个采集节点的类型标识，从对应的采集节点中获取采集到的数据集，该处理节点为主处理节点和接收到数据获取指令的从处理节点中的任一个，该数据集包括至少一条数据
402
任一处理节点从该多个处理节点共享的存储空间中获取异常检测模型，并通过异常检测模型对该数据集进行异常检测，确定该数据集中的异常数据
403
任一处理节点将该异常数据存储至该多个存储节点中的第一存储节点，该第一存储节点用于存储检测出的异常数据
404

图4

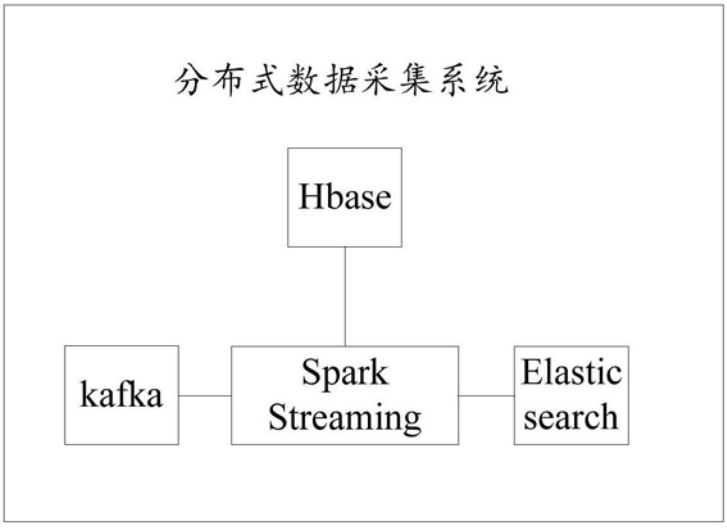

分布式数据采集系统
Hbase
kafka
Spark Streaming
Elastic search

图5

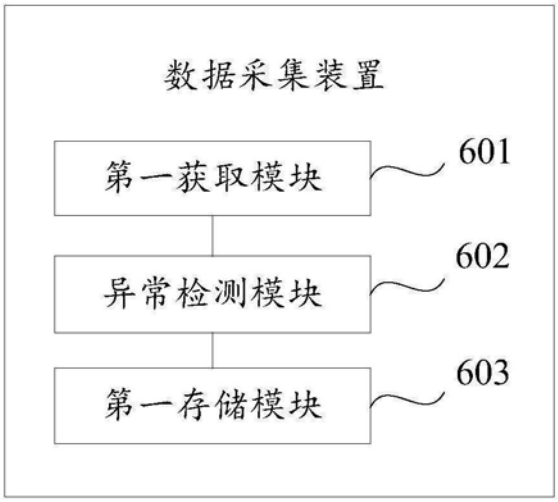

数据采集装置
第一获取模块
601
异常检测模块
602
第一存储模块
603

图6

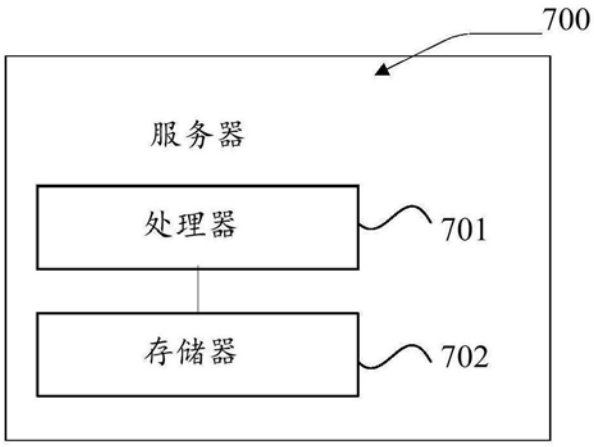

700
服务器
处理器
701
存储器
702

图7