(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115034278 A
(43)申请公布日 2022.09.09

(21)申请号 202110200046.8

(22)申请日 2021.02.22

(71)申请人 中国移动通信集团广东有限公司
地址 510623 广东省广州市天河区珠江新城珠江西路11号广东全球通大厦
申请人 中国移动通信集团有限公司

(72)发明人 叶芝高 何林艳 胡远明

(74)专利代理机构 北京路浩知识产权代理有限公司 11002
专利代理师 程琛

(51)Int.Cl.
G06K 9/62(2022.01)
G06F 11/34(2006.01)

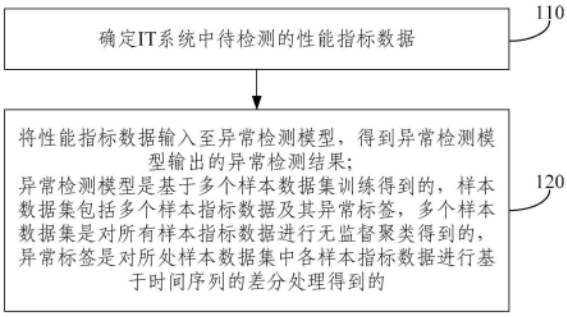
权利要求书2页 说明书10页 附图3页

(54)发明名称

性能指标异常检测方法、装置、电子设备和存储介质

(57)摘要

本发明提供一种性能指标异常检测方法、装置、电子设备和存储介质，其中方法包括：确定IT系统中待检测的性能指标数据；将性能指标数据输入至异常检测模型，得到异常检测模型输出的异常检测结果；异常检测模型基于多个样本数据集训练得到，样本数据集包括多个样本指标数据及其异常标签，多个样本数据集是对所有样本指标数据进行无监督聚类得到的，异常标签是对所处样本数据集中各样本指标数据进行基于时间序列的差分处理得到的。本发明提供的方法、装置、电子设备和存储介质，通过无监督聚类和基于时间序列的差分处理方式，实现了海量样本指标数据的自动标注，大大降低了异常检测的实现门槛，有助于提高性能指标数据异常检测的准确性和鲁棒性。



CN 115034278 A

1.一种性能指标异常检测方法,其特征在于,包括:

确定IT系统中待检测的性能指标数据;

将所述性能指标数据输入至异常检测模型,得到所述异常检测模型输出的异常检测结果;

所述异常检测模型是基于多个样本数据集训练得到的,所述样本数据集包括多个样本指标数据及其异常标签,所述多个样本数据集是对所有样本指标数据进行无监督聚类得到的,所述异常标签是对所处样本数据集中各样本指标数据进行基于时间序列的差分处理得到的。

2.根据权利要求1所述的性能指标异常检测方法,其特征在于,所述多个样本数据集的确定方法包括:

对所有样本指标数据进行无监督聚类,得到多个数据簇;

按照时序排列任一数据簇中各样本指标数据,得到所述任一数据簇中各指标的时间序列;

计算所述任一数据簇中各指标的时间序列在各基准时间点的突变指数,基于突变指数标记所述任一数据簇中各样本指标数据的异常标签,所述基准时间点是随机确定的;

基于所述任一数据簇中各样本指标数据及其异常标签,确定所述任一数据簇对应的样本数据集。

3.根据权利要求2所述的性能指标异常检测方法,其特征在于,所述计算所述任一数据簇中各指标的时间序列在各基准时间点的突变指数,包括:

基于基准时间点将所述时间序列切分为前后两个子序列;

基于所述前后两个子序列的均值和标准差,确定所述基准时间点的突变指数。

4.根据权利要求2所述的性能指标异常检测方法,其特征在于,所述对所有样本指标数据进行无监督聚类,得到多个数据簇,之前还包括:

基于主成分分析算法,对所有样本指标数据进行降维处理。

5.根据权利要求4所述的性能指标异常检测方法,其特征在于,所述基于主成分分析算法,对所有样本指标数据进行降维处理,包括:

结合奇异值分解算法和所述主成分分析算法,对所有样本指标数据进行降维处理。

6.根据权利要求2所述的性能指标异常检测方法,其特征在于,所述基于所述任一数据簇中各样本指标数据及其异常标签,确定所述任一数据簇对应的样本数据集,之后还包括:

将所述样本数据集中各样本指标数据的时间数据进行独热编码和/或标签编码。

7.根据权利要求1至6任一项所述的性能指标异常检测方法,其特征在于,所述异常检测模型是基于可扩展的大规模无监督离群点检测框架构建的。

8.一种性能指标异常检测装置,其特征在于,包括:

数据获取单元,用于确定IT系统中待检测的性能指标数据;

异常检测单元,用于将所述性能指标数据输入至异常检测模型,得到所述异常检测模型输出的异常检测结果;

所述异常检测模型是基于多个样本数据集训练得到的,所述样本数据集包括多个样本指标数据及其异常标签,所述多个样本数据集是对所有样本指标数据进行无监督聚类得到的,所述异常标签是对所处样本数据集中各样本指标数据进行基于时间序列的差分处理得

到的。

9.一种电子设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现如权利要求1至7任一项所述性能指标异常检测方法的步骤。

10.一种非暂态计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求1至7任一项所述性能指标异常检测方法的步骤。

性能指标异常检测方法、装置、电子设备和存储介质

技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种性能指标异常检测方法、装置、电子设备和存储介质。

背景技术

[0002] 目前基于IT系统性能指标的异常检测方案为一般分为单指标异常检测和多指标异常检测。其中，多指标异常检测即综合多个指标进行异常判断。

[0003] 针对于IT系统产生的高维数据，单指标异常检测结果可能会存在片面性，且单指标异常检测效率低，不切实际，而多指标异常检测则可以克服上述问题。多指标异常检测通常的实现方式为无监督学习算法和有监督分类算法两类。相较于无监督学习算法，有监督分类算法更加准确，鲁棒性更优，但是由于无法对海量的数据进行标注，有监督学习算法实现难度较高。

发明内容

[0004] 本发明提供一种性能指标异常检测方法、装置、电子设备和存储介质，用以解决现有的性能指标异常检测方法可靠性差、实现难度高的问题。

[0005] 本发明提供一种性能指标异常检测方法，包括：

[0006] 确定IT系统中待检测的性能指标数据；

[0007] 将所述性能指标数据输入至异常检测模型，得到所述异常检测模型输出的异常检测结果；

[0008] 所述异常检测模型是基于多个样本数据集训练得到的，所述样本数据集包括多个样本指标数据及其异常标签，所述多个样本数据集是对所有样本指标数据进行无监督聚类得到的，所述异常标签是对所处样本数据集中各样本指标数据进行基于时间序列的差分处理得到的。

[0009] 根据本发明提供的一种性能指标异常检测方法，对所有样本指标数据进行无监督聚类，得到多个数据簇；

[0010] 按照时序排列任一数据簇中各样本指标数据，得到所述任一数据簇中各指标的时间序列；

[0011] 计算所述任一数据簇中各指标的时间序列在各基准时间点的突变指数，基于突变指数标记所述任一数据簇中各样本指标数据的异常标签，所述基准时间点是随机确定的；

[0012] 基于所述任一数据簇中各样本指标数据及其异常标签，确定所述任一数据簇对应的样本数据集。

[0013] 根据本发明提供的一种性能指标异常检测方法，所述计算所述任一数据簇中各指标的时间序列在各基准时间点的突变指数，包括：

[0014] 基于基准时间点将所述时间序列切分为前后两个子序列；

[0015] 基于所述前后两个子序列的均值和标准差，确定所述基准时间点的突变指数。

[0016] 根据本发明提供的一种性能指标异常检测方法，所述对所有样本指标数据进行无监督聚类，得到多个数据簇，之前还包括：

[0017] 基于主成分分析算法，对所有样本指标数据进行降维处理。

[0018] 根据本发明提供的一种性能指标异常检测方法，所述基于主成分分析算法，对所有样本指标数据进行降维处理，包括：

[0019] 结合奇异值分解算法和所述主成分分析算法，对所有样本指标数据进行降维处理。

[0020] 根据本发明提供的一种性能指标异常检测方法，所述基于所述任一数据簇中各样本指标数据及其异常标签，确定所述任一数据簇对应的样本数据集，之后还包括：

[0021] 将所述样本数据集中各样本指标数据的时间数据进行独热编码和/或标签编码。

[0022] 根据本发明提供的一种性能指标异常检测方法，所述异常检测模型是基于可扩展的大规模无监督离群点检测框架构建的。

[0023] 本发明提供一种性能指标异常检测装置，包括：

[0024] 数据获取单元，用于确定IT系统中待检测的性能指标数据；

[0025] 异常检测单元，用于将所述性能指标数据输入至异常检测模型，得到所述异常检测模型输出的异常检测结果；

[0026] 所述异常检测模型是基于多个样本数据集训练得到的，所述样本数据集包括多个样本指标数据及其异常标签，所述多个样本数据集是对所有样本指标数据进行无监督聚类得到的，所述异常标签是对所处样本数据集中各样本指标数据进行基于时间序列的差分处理得到的。

[0027] 本发明还提供一种电子设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现如上述任一种所述性能指标异常检测方法的步骤。

[0028] 本发明还提供一种非暂态计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现如上述任一种所述性能指标异常检测方法的步骤。

[0029] 本发明提供的性能指标异常检测方法、装置、电子设备和存储介质，通过无监督聚类和基于时间序列的差分处理方式，实现了海量样本指标数据的自动标注，大大降低了异常检测的实现门槛，有助于提高性能指标数据异常检测的准确性和鲁棒性。

附图说明

[0030] 为了更清楚地说明本发明或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作一简单地介绍，显而易见地，下面描述中的附图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。

[0031] 图1是本发明提供的性能指标异常检测方法的流程示意图之一；

[0032] 图2是本发明提供的样本数据集确定方法的流程示意图之一；

[0033] 图3是本发明提供的性能指标异常检测方法的流程示意图之二；

[0034] 图4是本发明提供的样本数据集确定方法的流程示意图之二；

[0035] 图5是本发明提供的性能指标异常检测装置的结构示意图；

[0036] 图6是本发明提供的电子设备的结构示意图。

具体实施方式

[0037] 为使本发明的目的、技术方案和优点更加清楚，下面将结合本发明中的附图，对本发明中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0038] 异常检测也称为离群检测，用于发现数据流中的异常类型并确定其发生的详细信息。常见的异常检测分为单指标异常检测与多指标异常检测。

[0039] 其中，单指标异常检测即时间序列异常检测，通常为基于统计的算法。该算法非常简单和容易实现，但只能处理简单场景。随着整个互联网业务急剧膨胀，运行场景日趋复杂，IT系统指标数据规模会变大且各类监控指标亦会增加。面对大数据场景，传统的单指标异常检测结果可能会存在片面性，且检测效率低，不切实际。

[0040] 多指标异常检测即综合多个指标进行异常判断，具体可以分为无监督学习算法和有监督分类算法两类。无监督学习算法有IForest孤立森林、LOF局部异常因子、One-Class-SVM、Autoencoder自编码器等，这类算法无需对数据进行标注，但是特征选取难，检验结果准确性低，且方法执行的鲁棒性较低。有监督分类算法包括XGBoost、GBDT、决策树、支持向量机等，这类算法更为准确，但是标注的数据难以得到。有监督分类和无监督学习这两类多指标异常检测算法各有利弊，相较于无监督学习算法，有监督分类算法更加准确，鲁棒性更优，但是由于无法对海量的数据进行标注，有监督学习算法实现难度较高。

[0041] 针对上述问题，本发明实施例提供了一种性能指标异常检测方法。图1是本发明提供的性能指标异常检测方法的流程示意图之一，如图1所示，该方法包括：

[0042] 步骤110，确定IT系统中待检测的性能指标数据。

[0043] 此处，IT系统即需要进行性能指标异常检测的系统，待检测的性能指标数据可以包含IT系统中各个设备下各个接口产生的各个性能指标数据。

[0044] 步骤120，将性能指标数据输入至异常检测模型，得到异常检测模型输出的异常检测结果；

[0045] 异常检测模型是基于多个样本数据集训练得到的，样本数据集包括多个样本指标数据及其异常标签，多个样本数据集是对所有样本指标数据进行无监督聚类得到的，异常标签是对所处样本数据集中各样本指标数据进行基于时间序列的差分处理得到的。

[0046] 具体地，此处的异常检测模型是有监督分类得到的，其训练样本包括多个样本数据集，每个样本数据集中包含的样本指标数据均标注有异常标签，用于反映对应的样本指标数据是否存在异常。

[0047] 考虑到多指标异常检测模型训练所需的样本数据规模极其庞大，人工进行样本标注并不现实，本发明实施例通过无监督聚类和基于时间序列的差分处理方式，实现了样本数据的自动标注。

[0048] 多个样本数据集的获取可以通过如下步骤实现：首先，收集大量样本指标数据，并对所有样本指标数据进行无监督聚类，从而将所有样本指标数据划分为多个类，每一类样本指标数据对应一个样本数据集。分别对各类样本指标数据进行基于时间序列的差分处

理，从而快速定位各类样本指标数据中的异常数据，进而标注各样本指标数据的异常标签。

[0049] 在得到多个样本数据集之后，即可基于多个样本数据集进行模型训练，从而得到异常检测模型，经过有监督分类得到的异常检测模型，学习了样本指标数据和异常标签之间的映射关系，因此在步骤120中，可以直接将待检测的性能指标数据输入到异常检测模型，由异常检测模型应用学习所得的映射关系，将性能指标数据映射到其对应异常标签，即得到异常检测结果并输出。

[0050] 本发明实施例提供的方法，通过无监督聚类和基于时间序列的差分处理方式，实现了海量样本指标数据的自动标注，大大降低了异常检测的实现门槛，有助于提高性能指标数据异常检测的准确性和鲁棒性。

[0051] 基于上述实施例，图2是本发明提供的样本数据集确定方法的流程示意图之一，如图2所示，样本数据集的确定方法包括：

[0052] 步骤210，对所有样本指标数据进行无监督聚类，得到多个数据簇。

[0053] 此处，对所有样本指标数据进行无监督聚类的目的在于对不同周期类型的样本指标数据进行分类，从而为后续基于时间序列的差分处理做准备工作。

[0054] 步骤220，按照时序排列任一数据簇中各样本指标数据，得到该数据簇中各指标的时间序列；

[0055] 步骤230，计算该数据簇中各指标的时间序列在各基准时间点的突变指数，基于突变指数标记该数据簇中各样本指标数据的异常标签，基准时间点是随机确定的；

[0056] 步骤240，基于该数据簇中各样本指标数据及其异常标签，确定该数据簇对应的样本数据集。

[0057] 具体地，在完成无监督聚类之后，可以分别对各个数据簇执行步骤220-240，从而得到各个数据簇对应的样本数据集。此处以单一数据簇为例进行说明：

[0058] 数据簇中的各个样本指标数据均携带有其产生时间，可以按照产生时间的先后对数据簇中的各个样本指标数据进行排序，从而得到各个指标的时间序列。

[0059] 针对于时间序列，对其中突变量的记录即相对于对样本指标数据进行异常标记。因此在得到时间序列之后，即可对时间序列进行突变量检测。具体的检测方式可以是参数检验法、非参数检验法等，考虑到参数检验法对样本总体分布有相应要求，需要在序列样本分布符合检验方法的假设条件才可采用，作为优选，本发明实施例可以采用更加简单的非参数检验法。在运用非参数检验法时，不必预先设定序列样本的总体分布就能反应时间序列的内在结构和局部变化特征，分辨时间序列在不同尺度上的演变特征。

[0060] 针对于IT系统，可以通过随机种子设定，确定基准时间点，然后通过基准时间点将时间序列切割为前后两个子序列，基于切割所得的前后两个子序列计算基准时间点的突变指数，通过突变指数的大小判断基准时间点处的样本指标数据是否相对于时间序列中的其余样本指标数据存在异常，进而实现数据簇中各样本指标数据的异常标记。

[0061] 在完成异常标记后，即可基于数据簇中携带有异常标签的各样本指标数据构建该数据簇对应的样本数据集。

[0062] 基于上述任一实施例，步骤230包括：

[0063] 基于基准时间点将时间序列切分为前后两个子序列；

[0064] 基于前后两个子序列的均值和标准差，确定基准时间点的突变指数。

[0065] 具体地，针对于IT系统，可以通过信息和噪声两部分来讨论突变问题。具体在基于基准时间点切分开的前后两个子序列，并分别计算前后两个子序列的均值和标准差，其中均值可以代表信息，标准差可以代表噪声，由此可以确定基准时间点的突变指数$\frac{S}{N}$如下式所示：

[0066] $$\frac{S}{N} = |x_1 - x_2|/(S_1 + S_2)$$

[0067] 式中，x_1和S_1分别代表基准时间点之前m_1时间段的子序列的平均值和标准差，x_2和S_2分别代表基准时间点之后m_2时间段的子序列的平均值和标准差。

[0068] 通过在时间段中连续设置基准时间点，即可得到突变指数S/N的时间序列。在此基础上，可基于各突变指数S/N的大小，确定对应时间点处样本指标数据的异常标签。例如，当1<S/N<2的时候，将对应时间段处的样本指标数据的异常标签标记为“1”，S/N>2的时候异常标签标记为“2”。此处异常标签1可以表示正常，异常标签2可以表示异常。

[0069] 此外，实际生产环境中，IT系统有多个设备分别对应不同的指标数据。而不同的设备、接口，也会出现不同的指标数据。例如在数据库中，每个分区会有对应的容量消耗，数据库等待事件数等不同的指标数据。因此，IT系统产生的性能指标数据相当庞大，目前在对于高纬度的数据进行建模之前，都会进行降维处理。随着数据维度不断降低，数据存储所需的空间也会随之减少；低维数据有助于减少计算/训练用时；一些算法在高维度数据上容易表现不佳，降维可提高算法可用性等等。而现实中对于数据降维的处理方法较少，传统的降维方法耗时较久，计算成本也都很高。

[0070] 针对这一问题，基于上述任一实施例，在执行步骤210之前，还包括：

[0071] 步骤200，基于主成分分析算法，对所有样本指标数据进行降维处理。

[0072] 具体地，直接采集得到的样本指标数据存在较多维度，可以通过主成分分析算法(Principle Component Analysis，PCA)进行降维处理，从而降低后续计算量，提高后续建模效率。

[0073] 应用PCA进行降维的步骤如下：

[0074] 1、收集样本指标数据，假设收集所得的样本指标数据为m条n维；

[0075] 2、将样本指标数据按列组成n行m列的矩阵X；

[0076] 3、将矩阵X的每一行进行零均值化，即减去这一行的均值；

[0077] 4、求出协方差矩阵$C = \frac{1}{m}XX^T$；

[0078] 5、求出协方差矩阵的特征值及对应的特征向量；

[0079] 6、将特征向量按对应特征值大小从上到下按行排列成矩阵，取K行组成矩阵P；Y＝PX，即为降维到k维后的数据。

[0080] 基于上述任一实施例，步骤200包括：结合奇异值分解算法和主成分分析算法，对所有样本指标数据进行降维处理。

[0081] 具体地，IT系统获取所得的样本指标数据中，样本维度很高，PCA降维处理时的协方差矩阵计算太慢，且k保留的主要信息只是针对训练集的，主要信息未必是重要信息，有可能舍弃了一些看似无用但恰好是重要信息的信息。即单独应用PCA进行降维也可能会加剧过拟合。考虑到这一特殊情况，可以将PCA问题转换为SVD(Singular Value

Decomposition,奇异值分解)问题,从而避免XX^T的计算。

[0082] 进一步地,PCA求解的关键在于求解$C=\frac{1}{m}XX^T$的特征值分解,而SVD分解的关键在于A^TA的计算。可以取$A=\frac{X^T}{\sqrt{m}}$,则有:

[0083] $$A^TA=\left(\frac{X^T}{\sqrt{m}}\right)^T\frac{X^T}{\sqrt{m}}=\frac{1}{m}XX^T$$

[0084] 由此将SVD与PCA等价,以提高特征值迭代求解的效率,降低计算量。

[0085] 基于上述任一实施例,步骤240之后还包括:

[0086] 将样本数据集中各样本指标数据的时间数据进行独热编码和/或标签编码。

[0087] 具体地,针对各样本指标数据中携带的时间数据,即用于表征样本指标数据产生时间的信息,可以进行拆分。

[0088] 比如时间数据2020-10-08 10:30:00.360000+00:00,可以将其转换为日期类型,然后从里面提取年、月、日等时间信息,而针对其中的小时信息hour,也可以将小时信息hour按照时刻转换为离散数据,例如morning、afternoon等。针对于提取得到的各类别的时间信息,可以通过独热编码(one-hot encoding)和/或标签编码(label encoding)的方式进行编码。

[0089] 其中,独热编码可以通过逻辑代码编写完成。具备K个类别的变量的编码结果是一个K列的二值矩阵,其中第i列的值为1意味着这项观测属于第i类。标签编码直接将类别转换为数字,使用标签编码能够保持原本的维度。

[0090] 基于上述任一实施例,异常检测模型是基于可扩展的大规模无监督离群点检测框架构建的。

[0091] 具体地,可扩展的大规模无监督离群点检测框架(Scalable Unsupervised Outlier Detection Framework,SUOD)能够在加速训练的同时不牺牲训练和预测的效果,从而解决在高维大数据上训练多个异常检测模型无法收敛的问题。在异常检测模型的训练过程中,可以对各个样本数据集在SUOD框架中拟合多个有监督与无监督异常检测算法,从而低时高效地求出多个异常检测模型,以用于对待检测的性能指标数据进行预测判断。

[0092] 进一步地,SUOD框架可以包括三个模块,分别是随机降维、均衡并行调度和伪监督的模型训练分析。

[0093] 在随机降维方面:海量的数据模型拟合中,运算开销会随着维度上升而变大,而且往往是几何增长的,因此降维是一个非常合理的选择。但简单的降维方法,往往缺乏性能保证。而线性降维,比如PCA,其最大的问题是降维所得的结果是确定的,也就是说每次通过PCA得到的结果都是差不多的,但这并不利于后续多个模型的训练,尤其不利于集成学习模型的训练。另一个显著的问题就是PCA本身就可以被作为一种异常检测的方法,降维的过程会丢失掉一部分异常点。

[0094] 针对这个问题,在数据层面上,可以通过Johnson-Lindenstrauss(JL)projection满足降维需求。JL实际操作非常简单,只要随机出一个投影矩阵(比如完全随机的高斯分布)即可。

[0095] 在系统均衡调度方面:并行训练多个异质模型,会导致很大的运行开销差别。例

如,决策树训练可能比K近邻的运算速度要快的多。假设训练100个模型,前50个是决策树,后50个是K近邻。简单将前50个和后50个划分为2个集群或者进程,那么集群1上的任务会比集群2上的任务完成速度快的多。这种调度上的不平衡会使得整个系统的运行效率很低。

[0096] 为了解决异质模型在系统调度上的不平衡,可以首先训练一个回归模型,用来预测每个学习器所需的运算开销。随后基于各学习器所需的运算开销,设定调度方法来确保不同的上的任务负载尽量接近。通过调度方法的设置和运行,所得训练过程中每个进程上的任务数量可能不同,但总的需要时间会更加接近,因此并行任务可以在几乎同一时刻完成。

[0097] 在模型训练方面:无监督模型,尤其是非参数的无监督模型,往往预测的开销很大。例如K近邻的预测就比较耗时。大部分无监督的异常检测模型的预测开销都很大。考虑到这个问题,可以选择参数化的监督模型来替换无监督模型,即通过Pseudo supervised approximation(伪监督拟合,PSA)的思路,在已有数据上用无监督异常检测来训练,并得到它们的训练分数,再用监督学习模型来模拟决策边界。

[0098] 图3是本发明提供的性能指标异常检测方法的流程示意图之二,如图3所示,性能指标异常检测方法包括:

[0099] S1,样本数据集处理:

[0100] 将一段时间内,IT系统的各个关键设备的指标数据进行收集;作为样本指标数据。在此基础上,通过设备关联关系,以及PCA方法将指标口径统一,压缩并去除无关指标数据。随后通过无监督聚类算法,将不同周期类型的时序数据进行分类,并通过一阶差分后的新变量生成和时间日期的one-hot编码,最后获取多个处理后的样本数据集。

[0101] S2,异常检测模型训练:

[0102] 基于步骤S1获取到的多个样本数据集,在SUOD框架中对步骤S1获取到的多个样本数据集进行无监督或伪监督的异常检测算法拟合,从而得到包含多种异常检测算法的异常检测模型。

[0103] S3,异常检测:

[0104] 在得到待检测的性能指标数据后,可以将性能指标数据输入至步骤S2所得的异常检测模型,得到异常检测模型输出的异常检测结果。在此基础上,还可以将异常检测结果在WEB/APP上呈现,方便运维人员查询使用,以提高对云原生架构的IT系统异常检测的智能化、效率化。

[0105] 基于上述任一实施例,图4是本发明提供的样本数据集确定方法的流程示意图之二,如图4所示,S1包括:

[0106] S1.1,收集一段时间内IT系统的各个关键设备的指标数据,作为样本指标数据。

[0107] S1.2,采用PCA,或者采用基于SVD进行改进后的PCA对样本指标数据进行降维。

[0108] S1.3,对降维后的所有样本指标数据进行无监督聚类,得到多个数据簇。

[0109] S1.4,将各数据簇中的样本指标数据按照时序排列,进而在时间序列的基础上进行数据差分计算,从而实现各样本指标数据的异常标记。

[0110] S1.5,将各样本指标数据的时间数据进行独热编码和/或标签编码。

[0111] S1.6,结合各数据簇中的各样本指标数据的异常标记和时间编码,构建各数据簇分别对应的样本数据集。

[0112] 本发明实施例针对IT系统的高维多指标数据，提出了一种数据处理、模型框架分析的新思路，通过PCA算法降维，实现了更高效率更有效的数据降维和降噪；通过无监督聚类，更好地对数据进行划分。通过基于时间序列的差分处理进行数据标注，相比简单融合特征变量更能反映出数据的变换趋势。由此得到的多个样本数据集，为模型拟合做出了更好的准备工作。

[0113] 此外，本发明实施例通过SUOD模型集成框架对分类数据进行多个模型的建立，多模型的异常检测方案，将数据集划分为不同类型，本质上是利用到了时间序列的周期性和趋势性，而SUOD框架不仅仅是训练多个模型，而是通过数据压缩，均衡调度等多个方面，提升模型训练的速度，并且不会降低模型检测的效果。最后通过模型集合，输入相关检测数据就可以检测出IT系统中的异常时间点以及对应的指标值。因此运维人员能更加方便的发现与验证IT系统中的异常点，并准确地找到相关的异常点和根因指标。

[0114] 基于上述任一实施例，图5是本发明提供的性能指标异常检测装置的结构示意图，如图5所示，该装置包括：

[0115] 数据获取单元510，用于确定IT系统中待检测的性能指标数据；

[0116] 异常检测单元520，用于将所述性能指标数据输入至异常检测模型，得到所述异常检测模型输出的异常检测结果；

[0117] 所述异常检测模型是基于多个样本数据集训练得到的，所述样本数据集包括多个样本指标数据及其异常标签，所述多个样本数据集是对所有样本指标数据进行无监督聚类得到的，所述异常标签是对所处样本数据集中各样本指标数据进行基于时间序列的差分处理得到的。

[0118] 本发明实施例提供的装置，通过无监督聚类和基于时间序列的差分处理方式，实现了海量样本指标数据的自动标注，大大降低了异常检测的实现门槛，有助于提高性能指标数据异常检测的准确性和鲁棒性。

[0119] 基于上述任一实施例，该装置还包括样本确定单元，用于：

[0120] 对所有样本指标数据进行无监督聚类，得到多个数据簇；

[0121] 按照时序排列任一数据簇中各样本指标数据，得到所述任一数据簇中各指标的时间序列；

[0122] 计算所述任一数据簇中各指标的时间序列在各基准时间点的突变指数，基于突变指数标记所述任一数据簇中各样本指标数据的异常标签，所述基准时间点是随机确定的；

[0123] 基于所述任一数据簇中各样本指标数据及其异常标签，确定所述任一数据簇对应的样本数据集。

[0124] 基于上述任一实施例，样本确定单元用于：

[0125] 基于基准时间点将所述时间序列切分为前后两个子序列；

[0126] 基于所述前后两个子序列的均值和标准差，确定所述基准时间点的突变指数。

[0127] 基于上述任一实施例，样本确定单元还用于：

[0128] 基于主成分分析算法，对所有样本指标数据进行降维处理。

[0129] 基于上述任一实施例，样本确定单元还用于：

[0130] 结合奇异值分解算法和所述主成分分析算法，对所有样本指标数据进行降维处理。

[0131] 基于上述任一实施例,样本确定单元还用于:

[0132] 将所述样本数据集中各样本指标数据的时间数据进行独热编码和/或标签编码。

[0133] 基于上述任一实施例,所述异常检测模型是基于可扩展的大规模无监督离群点检测框架构建的。

[0134] 图6示例了一种电子设备的实体结构示意图,如图6所示,该电子设备可以包括:处理器(processor)610、通信接口(Communications Interface)620、存储器(memory)630和通信总线640,其中,处理器610,通信接口620,存储器630通过通信总线640完成相互间的通信。处理器610可以调用存储器630中的逻辑指令,以执行性能指标异常检测方法,该方法包括:确定IT系统中待检测的性能指标数据;将所述性能指标数据输入至异常检测模型,得到所述异常检测模型输出的异常检测结果;所述异常检测模型是基于多个样本数据集训练得到的,所述样本数据集包括多个样本指标数据及其异常标签,所述多个样本数据集是对所有样本指标数据进行无监督聚类得到的,所述异常标签是对所处样本数据集中各样本指标数据进行基于时间序列的差分处理得到的。

[0135] 此外,上述的存储器630中的逻辑指令可以通过软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0136] 另一方面,本发明还提供一种计算机程序产品,所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算机程序,所述计算机程序包括程序指令,当所述程序指令被计算机执行时,计算机能够执行上述各方法所提供的性能指标异常检测方法,该方法包括:确定IT系统中待检测的性能指标数据;将所述性能指标数据输入至异常检测模型,得到所述异常检测模型输出的异常检测结果;所述异常检测模型是基于多个样本数据集训练得到的,所述样本数据集包括多个样本指标数据及其异常标签,所述多个样本数据集是对所有样本指标数据进行无监督聚类得到的,所述异常标签是对所处样本数据集中各样本指标数据进行基于时间序列的差分处理得到的。

[0137] 又一方面,本发明还提供一种非暂态计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现以执行上述各提供的性能指标异常检测方法,该方法包括:确定IT系统中待检测的性能指标数据;将所述性能指标数据输入至异常检测模型,得到所述异常检测模型输出的异常检测结果;所述异常检测模型是基于多个样本数据集训练得到的,所述样本数据集包括多个样本指标数据及其异常标签,所述多个样本数据集是对所有样本指标数据进行无监督聚类得到的,所述异常标签是对所处样本数据集中各样本指标数据进行基于时间序列的差分处理得到的。

[0138] 以上所描述的装置实施例仅仅是示意性的,其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选取其

中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下,即可以理解并实施。

[0139] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件。基于这样的理解,上述技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品可以存储在计算机可读存储介质中,如ROM/RAM、磁碟、光盘等,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0140] 最后应说明的是:以上实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

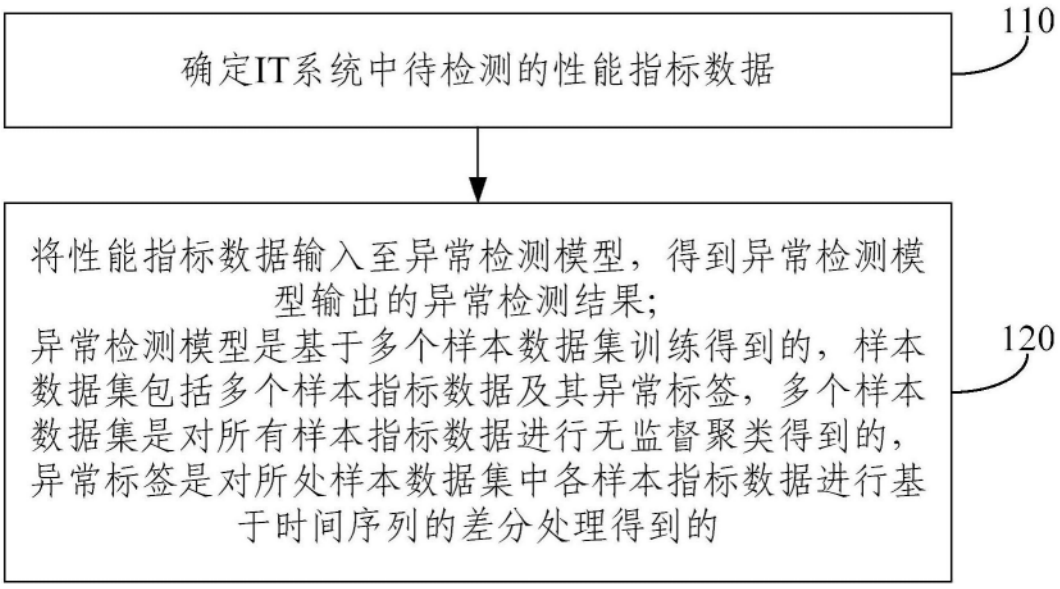

确定IT系统中待检测的性能指标数据
110
将性能指标数据输入至异常检测模型，得到异常检测模型输出的异常检测结果;
异常检测模型是基于多个样本数据集训练得到的，样本数据集包括多个样本指标数据及其异常标签，多个样本数据集是对所有样本指标数据进行无监督聚类得到的，异常标签是对所处样本数据集中各样本指标数据进行基于时间序列的差分处理得到的
120

图1

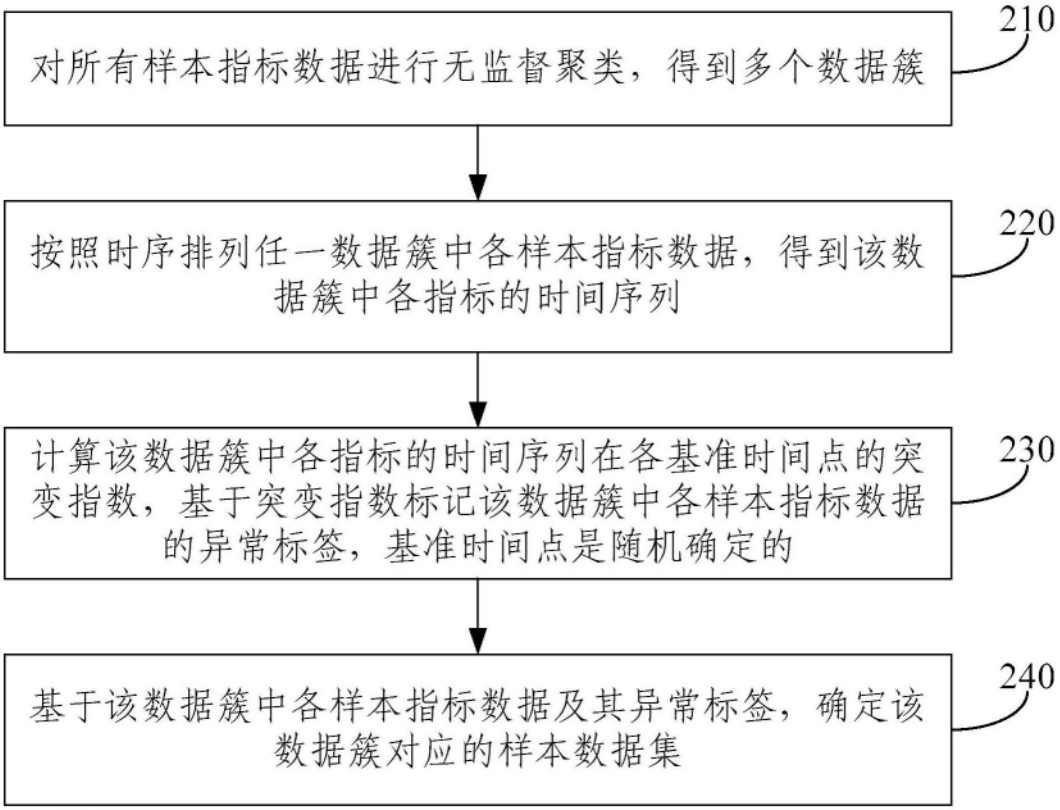

对所有样本指标数据进行无监督聚类，得到多个数据簇
210
按照时序排列任一数据簇中各样本指标数据，得到该数据簇中各指标的时间序列
220
计算该数据簇中各指标的时间序列在各基准时间点的突变指数，基于突变指数标记该数据簇中各样本指标数据的异常标签，基准时间点是随机确定的
230
基于该数据簇中各样本指标数据及其异常标签，确定该数据簇对应的样本数据集
240

图2

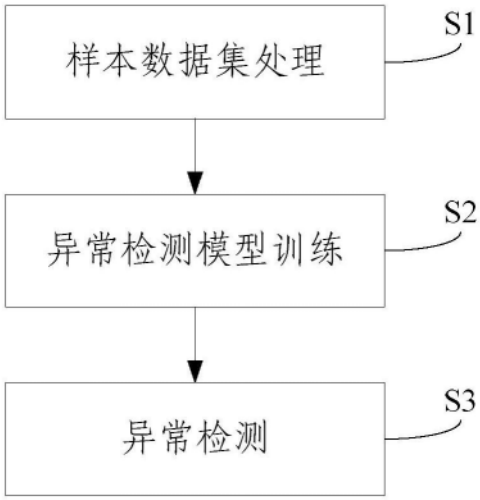

样本数据集处理
S1
异常检测模型训练
S2
异常检测
S3

图3

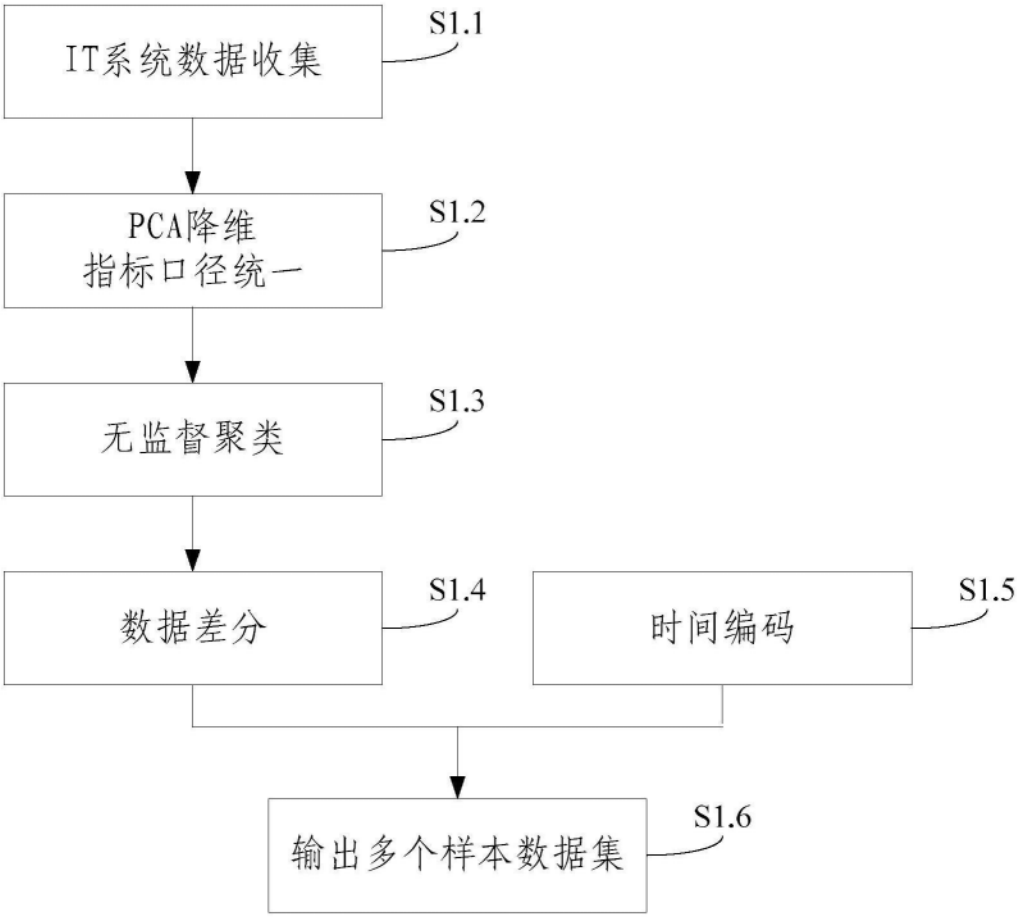

IT系统数据收集
S1.1
PCA降维
指标口径统一
S1.2
无监督聚类
S1.3
数据差分
S1.4
时间编码
S1.5
输出多个样本数据集
S1.6

图4

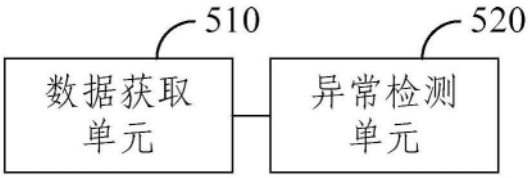


图5

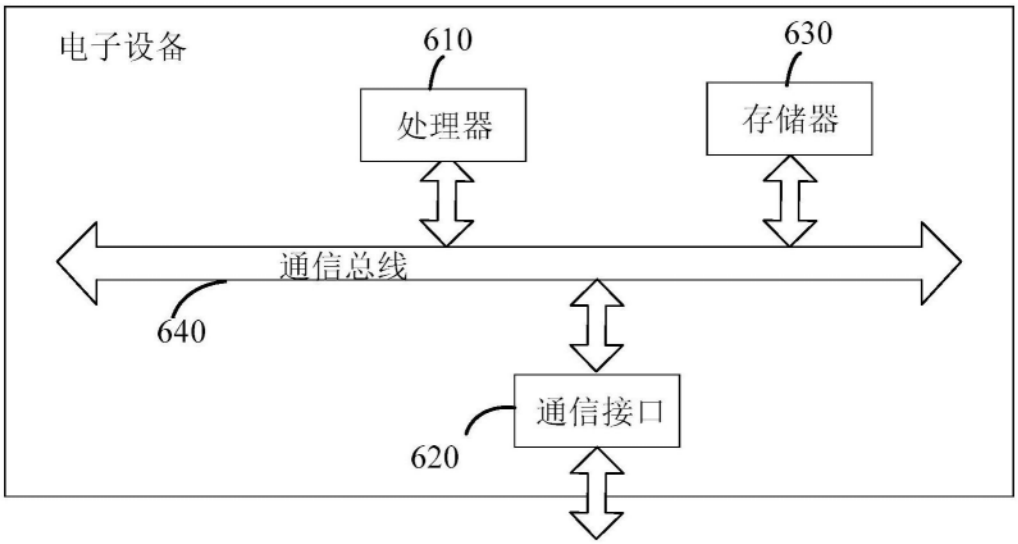


图6