(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115964248 A
(43)申请公布日 2023.04.14

(21)申请号 202211705524.1

(22)申请日 2022.12.29

(71)申请人 南京大学
地址 210023 江苏省南京市栖霞区仙林大道163号

(72)发明人 顾胜晖 李显 荣国平 张贺 邵栋 周鑫

(74)专利代理机构 南京众联专利代理有限公司 32206
专利代理师 顾进

(51)Int.Cl.
G06F 11/30(2006.01)
G06F 16/18(2019.01)
G06N 3/045(2023.01)
G06N 3/08(2023.01)

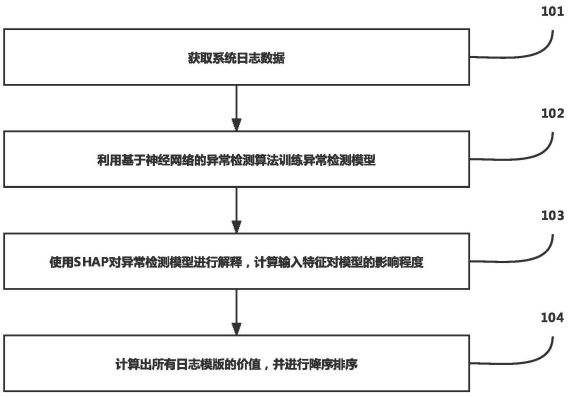
权利要求书2页 说明书6页 附图2页

(54)发明名称

系统日志评价方法、装置、设备及介质

(57)摘要

本发明涉及系统日志评价方法，所述方法包括以下步骤：步骤1：获取系统日志数据，步骤2：提取系统日志数据模板，将非结构化的系统日志数据转化为结构化的系统日志，步骤3：利用基于神经网络的异常检测算法训练异常检测模型，步骤4：对异常检测模型进行解释，输出系统日志价值。该技术方案通过提高日志信息量，减少冗余日志等方式，提升日志数据的质量，从而达到提升数据价值的同时减少数据收集量的效果，进一步提升异常检测算法的效果与效率。



CN 115964248 A

1.系统日志评价方法,其特征在于,所述方法包括以下步骤:

步骤1:获取系统日志数据,

步骤2:提取系统日志数据模板,将非结构化的系统日志数据转化为结构化的系统日志,

步骤3:利用基于神经网络的异常检测算法训练异常检测模型,

步骤4:对异常检测模型进行解释,输出系统日志价值。

2.根据权利要求1所述的系统日志评价方法,其特征在于,步骤1中获取系统日志数据,包括:获取系统日志内容,包括时间戳、服务标识符、服务所在主机标志符、日志事件类型、日志信息、执行次数以及事件持续时长。

3.根据权利要求1所述的系统日志评价方法,其特征在于,步骤2中包括:将系统日志数据输入至日志解析工具中,根据日志解析工具的输出结果提取日志数据模版;根据日志数据模版将非结构化的系统日志数据处理成为结构化的系统日志。

4.根据权利要求1所述的系统日志评价方法,其特征在于,步骤3中,利用基于神经网络的异常检测算法训练异常检测模型,包括:将结构化系统日志作为输入,将日志分组后,转化为异常检测算法所需的日志事件向量;利用预设的基于神经网络的异常检测算法,定位异常日志事件序列,构建异常检测模型;计算异常检测模型在验证集的F1值、准确率、精确率和召回率。

5.根据权利要求4所述的系统日志评价方法,其特征在于,步骤3中的日志分组,包括:使用日志时间戳和日志标识符,将日志划分为不同的组,每个组代表一个日志序列,使用固定窗口分组,具有预定义的窗口大小,指示用于拆分计时排序日志的时间间隔,两个连续的固定窗口之间没有重叠,出现在固定窗口中的所有日志信息将被分组到同一个日志序列中,使用滑动窗口分组,具有预定义的窗口大小和步长,步长一般小于窗口大小,使不同滑动窗口之间的重叠,相邻的日志序列中包含相同的日志信息,使用会话窗口分组,根据在日志信息中的唯一标识进行分组,同一个窗口的认知信息有相同的唯一标识,不同的会话窗口的唯一标识不同,步骤3中预设基于神经网络的异常检测算法包括DeepLog算法、RobustLog算法、Logsy算法、CNN算法和Autoencoder算法。

6.根据权利要求4所述的系统日志评价方法,其特征在于,步骤3中日志事件向量,包括:序列向量,反映了日志信息在窗口中的顺序;

数量向量,表示窗口中每个日志模板出现的次数;

语义向量,来自于语言模型,以表示日志信息的语义。

7.根据权利要求4所述的系统日志评价方法,其特征在于,步骤4中,对异常检测模型进行解释,输出系统日志价值,包括:将异常检测模型作为输入,将训练模型所使用的训练数据集作为背景数据;通过随机使用背景数据替换特征值的方式,计算出被遮盖特征的近似Shapley值;计算所有日志模板的对异常检测结果的平均贡献值,生成日志模板的价值排序。

8.一种系统日志价值评价装置,其特征在于,该装置用于实现权利要求1-7任意一项系统日志评价方法,所述装置包括:

日志收集模块,用于采集存储在不同位置的日志数据;

日志解析模块,用于将非结构化日志文件解析为日志模板和结构化文件;

数据预处理模块,用于将结构化日志处理为向量数据;

异常检测模型训练模块,用于对日志异常检测模型进行训练以及评估,输出训练完成的模型;异常检测模型解释模块,用于解释异常检测模型,输出系统日志模版的价值排序。

9.一种系统日志价值评价设备,其特征在于,所述设备包括:

一个或多个处理器;

存储装置,用于存储一个或多个程序;

当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现如权利要求1-8中任一所述的系统日志价值评价方法。

10.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,该程序被处理器执行时实现如权利要求1-8中任一所述系统日志价值评价方法。

系统日志评价方法、装置、设备及介质

技术领域

[0001] 本发明实施例涉及数据处理技术领域，尤其涉及一种系统日志价值评价方法、装置、设备及介质。

背景技术

[0002] 如今，我们日常生活中提供各种服务的软件系统越来越普遍，如搜索引擎、社交媒体、翻译应用程序等。与传统的单机软件不同，在线软件通常为全球数亿客户提供服务，目标是全天候可用性，如何保障服务质量成为市场上的核心竞争力。日志一直是确保许多软件系统的可靠性和连续性的必要数据，记录系统运行时的信息，促进系统故障排除和行为理解。现代软件系统日趋复杂，日志数量已经达到了前所未有的水平。

[0003] 近年来，随着服务监控和可观测性的重要性越来越被从业者所认可，大量的研究致力于运用人工智能技术对各类监控数据进行处理分析，以实现异常检测或分析等运维任务。而监控数据的质量是上述异常检测或分析任务的基石。在众多类型的监控数据中，日志数据的质量往往是异常检测效果的关键，其非结构化的特性以及不恰当的埋点极易导致关键信息的缺失以及产生大量低价值的冗余数据。

[0004] 日志数据中易缺失重要的信息。大部分异常检测和分析方法依赖日志数据中的关键信息进行决策。由于日志语句主要依赖开发者的项目经验以及偏好，因此日志语句的内容难免会缺失部分重要信息或变量。而重要的日志信息的缺失会导致异常检测和分析系统的准确度下降。日志数据中存在大量冗余和无效的信息。随着微服务系统体量与复杂度的增加，日志数据成爆发式增长。对于特定的异常检测和分析任务而言，并不是所有的日志信息都能提供价值，这些低价值的日志数据会影响异常检测和分析系统的效果与性能。筛选出高价值的日志信息有助于提高异常检测和分析任务的性能。

发明内容

[0005] 本发明针对针对现有技术中存在的技术问题，提供一种系统日志评价方法、装置、设备及介质，以合理、有效地对系统日志进行价值评估，形成系统日志评价模型，用于筛选日志数据。

[0006] 为了实现上述目的，本发明的技术方案如下：系统日志评价方法，所述方法包括以下步骤：

[0007] 步骤1：获取系统日志数据，

[0008] 步骤2：提取系统日志数据模板，将非结构化的系统日志数据转化为结构化的系统日志，

[0009] 步骤3：利用基于神经网络的异常检测算法训练异常检测模型，

[0010] 步骤4：对异常检测模型进行解释，输出系统日志价值。

[0011] 其中，步骤1中获取系统日志数据，包括：获取系统日志内容，包括但不限于时间戳、服务标识符、服务所在主机标志符、日志事件类型、日志信息、执行次数以及事件持续时

长。

[0012] 步骤2中包括:将系统日志数据输入至日志解析工具中,根据日志解析工具的输出结果提取日志数据模版;根据日志数据模版将非结构化的系统日志数据处理成为结构化的系统日志。其中,步骤3中,利用基于神经网络的异常检测算法训练异常检测模型,包括:将结构化系统日志作为输入,将日志分组后,转化为异常检测算法所需的日志事件向量;利用预设的基于神经网络的异常检测算法,定位异常日志事件序列,构建异常检测模型;计算异常检测模型在验证集的F1值、准确率、精确率和召回率。

[0013] 步骤3中的日志分组,包括:

[0014] 使用日志时间戳和日志标识符,将日志划分为不同的组,每个组代表一个日志序列,使用固定窗口分组,具有预定义的窗口大小,指示用于拆分计时排序日志的时间间隔,两个连续的固定窗口之间没有重叠,出现在固定窗口中的所有日志信息将被分组到同一个日志序列中。使用滑动窗口分组,具有预定义的窗口大小和步长,步长一般小于窗口大小,使不同滑动窗口之间的重叠,相邻的日志序列中包含相同的日志信息,使用会话窗口分组,根据在日志信息中的唯一标识进行分组,同一个窗口的认知信息有相同的唯一标识,不同的会话窗口的唯一标识不同。

[0015] 步骤3中预设基于神经网络的异常检测算法包括DeepLog算法、RobustLog算法、Logsy算法、CNN算法和Autoencoder算法。

[0016] 步骤3中日志事件向量,包括:

[0017] 序列向量,反映了日志信息在窗口中的顺序;

[0018] 数量向量,表示窗口中每个日志模板出现的次数;

[0019] 语义向量,来自于语言模型,以表示日志信息的语义。

[0020] 步骤4中,对异常检测模型进行解释,输出系统日志价值,包括:将异常检测模型作为输入,将训练模型所使用的训练数据集作为背景数据;通过随机使用背景数据替换特征值的方式,计算出被遮盖特征的近似Shapley值;计算所有日志模板的对异常检测结果的平均贡献值,生成日志模板的价值排序。

[0021] 一种系统日志价值评价装置,所述装置包括:

[0022] 日志收集模块,用于采集存储在不同位置的日志数据;

[0023] 日志解析模块,用于将非结构化日志文件解析为日志模板和结构化文件;

[0024] 数据预处理模块,用于将结构化日志处理为向量数据;

[0025] 异常检测模型训练模块,用于对日志异常检测模型进行训练以及评估,输出训练完成的模型;异常检测模型解释模块,用于解释异常检测模型,输出系统日志模版的价值排序。

[0026] 一种系统日志价值评价设备,所述设备包括:

[0027] 一个或多个处理器;

[0028] 存储装置,用于存储一个或多个程序;

[0029] 当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现如权利要求1-8中任一所述的系统日志价值评价方法。

[0030] 一种计算机可读存储介质,其上存储有计算机程序,其特征在于,该程序被处理器执行时实现如权利要求1-8中任一所述系统日志价值评价方法。

[0031] 相对于现有技术，本发明的优点如下：本发明实施例的技术方案，收集系统日志，将其转化为日志模版和结构化数据，经预处理为向量数据后训练异常检测模型，通过解释日志异常检测模型以及分析异常检测结果，对日志数据的质量进行价值评估，识别其中价值较低或者缺失的日志数据。通过提高日志信息量，减少冗余日志等方式，提升日志数据的质量，从而达到提升数据价值的同时减少数据收集量的效果，进一步提升异常检测算法的效果与效率。

附图说明

[0032] 图1是本发明实施例一提供的一种系统日志价值评价方法的流程图；

[0033] 图2是本发明实施例三中的一种系统日志价值评价装置的结构示意图；

[0034] 图3是本发明实施例四提供的一种设备的结构示意图。

具体实施方式

[0035] 为了加深对本发明的理解，下面结合附图对本实施例做详细的说明。

[0036] 实施例1：参见图1，

[0037] 图1为本发明实施例一中的一种系统日志价值评价方法的流程图，本实施例的技术方案适用于分布式系统下评估日志价值的情况，该方法可以由分布式系统日志价值评价装置执行，该装置可以由软件和/或硬件来实现，并可以集成在各种通用计算机设备中，具体包括如下步骤：

[0038] 步骤101、获取系统日志数据。

[0039] 本实施例中，采集了分布在不同位置的系统日志，包括时间戳、服务标识符、服务所在主机标志符、日志事件类型、日志信息、执行次数、事件持续时长等信息。

[0040] 步骤102、提取系统日志数据模板，将非结构化的系统日志数据转化为结构化的系统日志本实施例中，提取上一步获取的日志数据模版，将非结构化的系统日志数据转化为结构化的系统日志。具体的，日志数据通常是以非结构化或半结构化的形式存储在文件中的，基于启发式的日记解析算法相比其他来行的算法在准确率和时间效率上有更好的性能。本实施例使用Drain算法(Pinjia He,Jieming Zhu,Zibin Zheng,and Michael R.Lyu.2017.Drain:An Online Log ParsingApproach with Fixed Depth Tree.In 2017 IEEE International Conference on Web Services(ICWS)(ICWS'17).https://doi.org/10.1109/icws.2017.13)对上一步骤获取的原始日志数据进行解析，生成机构化的日志信息，并提取日志模版。

[0041] 步骤103、利用基于神经网络的异常检测算法训练异常检测模型

[0042] 本实施例中，将日志信息按照一个滑动窗口的方式进行分组，形成日志序列，转换成语义向量，使用预处理之后的日志数据对预置的日志异常检测模型进行训练及评估，最终输出训练完成的模型。

[0043] 步骤104、使用SHAP对异常检测模型进行解释，输出系统日志价值本实施例中，使用SHAP对训练后的模型进行解释，即计算异常检测模型输入的特征对模型结果的影响程度。具体的，本实施例将训练模型所使用的训练数据集作为背景数据，随机取样数据代替被遮盖的特征，计算出被遮盖特征的近似Shapely值，也就是该特征对检测结果的贡献值。日

志异常检测模型输入的特征向量中每一个维度的值代表的是日志模板的编号，且不同特征向量中的日志模板及其对应的序列大部分情况下并不相同，先将特征值根据日志模板编号顺序进行重新排序，然后再进行统计。计算日志模板的对异常检测结果的平均贡献值，结合日志模版对应的日志数量，输出日志模版的价值排序，具体计算公式如下：

[0044] $$R_j = \frac{|\phi_j|}{\sum_{i=1}^{n}|\phi_i|} + \frac{\log_2 \frac{c_j}{\sum_{i=1}^{n} c_i}}{\sum_{k=1}^{n} \log_2 \frac{c_k}{\sum_{i=1}^{n} c_i}}$$

[0045] 其中，R_j代表该日志模版对应价值权重；

[0046] Φ_j代表该日志模版的近似Shapely值；

[0047] c_j代表该日志模版的日志数量。

[0048] 实施例2：

[0049] 图2为本发明实施例二提供的一种系统日志价值评价装置的结构示意图，该日志价值评价装置，包括：日志收集模块201、日志解析模块202、数据预处理模块203、异常检测模型训练模块204、异常检测模型解释模块205。

[0050] 日志收集模块201，用于采集存储在不同位置的日志数据；

[0051] 日志解析模块202，用于将非结构化日志文件解析为日志模板和结构化文件；

[0052] 数据预处理模块203，用于将结构化日志处理为向量数据；

[0053] 异常检测模型训练模块204，用于训练异常检测模型并计算模型的指标；

[0054] 异常检测模型解释模块205，用于解释异常检测模型，输出系统日志模版的价值排序。

[0055] 本实施例的技术方案，获取系统日志数据；提取系统日志数据模板，将非结构化的系统日志数据转化为结构化的系统日志；利用基于神经网络的异常检测算法训练异常检测模型；使用SHAP对异常检测模型进行解释，输出系统日志价值，即本实施例的技术方案中，通过采集存储在不同位置的日志数据，将其转化为结构化日志数据，处理为向量数据训练异常检测模型后，经SHAP模型解释，能够合理、有效地对系统日志进行价值评估，形成系统日志评价模型，用于筛选日志数据，进而达到提升数据价值的同时减少数据收集量的目的。

[0056] 可选的，所述数据预处理模块，包括：

[0057] 固定窗口单元，使用固定窗口分组，预定义窗口大小，使用日志时间戳和日志标识符，将日志划分为不同的组，每个组代表一个日志序列，两个连续的固定窗口之间没有重叠。

[0058] 滑动窗口单元，使用滑动窗口分组，预定义窗口大小和步长大小，步长一般小于窗口大小，使用日志时间戳和日志标识符，将日志划分为不同的组，每个组代表一个日志序列，不同窗口分区之间的重叠。

[0059] 会话窗口单元，根据在日志信息中的唯一标识进行分组，同一个窗口的认知信息有相同的唯一标识，不同的会话窗口的唯一标识不同。

[0060] 可选的，所述数据预处理模块，包括：

[0061] 序列向量数据预处理单元，序列向量反映了日志信息在窗口中的顺序；

[0062] 数量向量数据预处理单元，数量向量表示窗口中每个日志模板出现的次数；

[0063] 语义向量数据预处理单元，语义向量来自于语言模型，以表示日志信息的语义。

[0064] 可选的，所述异常检测模型训练模块，包括：LSTM算法单元、CNN算法单元、Transformer算法单元和Autoencoder算法单元，用于满足不同的异常检测模型训练需求。

[0065] 本发明实施例所提供的系统日志价值评价装置可执行本发明任意实施例所提供的系统日志价值评价方法，具备执行方法相应的功能模块和有益效果。

[0066] 实施例3

[0067] 图3为本发明实施例三提供的一种设备的结构示意图，如图3所示，本申请中的计算机设备可以包括：

[0068] 一个或多个处理器31和存储器32；该计算机设备的处理器31可以是一个或多个，图3中以一个处理器31为例；存储器32用于存储一个或多个程序；所述一个或多个程序被所述一个或多个处理器31执行。

[0069] 计算机设备中的处理器31、存储器32可以通过总线或其他方式连接，图3中以通过总线连接为例。

[0070] 存储器32作为一种计算机可读存储介质，可设置为存储软件程序、计算机可执行程序以及模块。存储器32可包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序；存储数据区可存储根据设备的使用所创建的数据等。此外，存储器32可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。在一些实例中，存储器32可进一步包括相对于处理器31远程设置的存储器，这些远程存储器可以通过网络连接至计算机设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0071] 处理器31通过运行存储在存储器32中的程序，从而执行各种功能应用以及数据处理，例如实现本发明上述实施例所提供的系统日志价值评价方法。

[0072] 也即，所述处理单元执行所述程序时实现：获取系统日志数据；提取系统日志数据模板，将非结构化的系统日志数据转化为结构化的系统日志；利用基于神经网络的异常检测算法训练异常检测模型；使用SHAP对异常检测模型进行解释，输出系统日志价值。

[0073] 实施例4

[0074] 本发明实施例四还提供一种包含计算机可执行指令的存储介质，所述计算机可执行指令在由计算机处理器执行时用于执行一种系统日志价值评价方法，该方法包括：

[0075] 获取系统日志数据；提取系统日志数据模板，将非结构化的系统日志数据转化为结构化的系统日志；利用基于神经网络的异常检测算法训练异常检测模型；使用SHAP对异常检测模型进行解释，输出系统日志价值。

[0076] 当然，本发明实施例所提供的包含计算机可执行指令的存储介质，其计算机可执行指令不限于如上所述的方法操作，还可以执行本发明任意实施例所提供的系统日志价值评价方法中的相关操作。

[0077] 通过以上关于实施方式的描述，所属领域的技术人员可以清楚地了解到，本发明可借助软件及必需的通用硬件来实现，当然也可以通过硬件实现，但很多情况下前者是更佳的实施方式。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，如计算机的软盘、只读存储器(Read-Only Memory，ROM)、随机存取存储器(Random Access Memory，RAM)、闪存(FLASH)、硬盘或光盘等，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)执行本发明各个实施例所述的方法。

[0078] 可以以一种或多种程序设计语言或其组合来编写用于执行本发明操作的计算机程序代码,所述程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++,还包括常规的过程式程序设计语言——诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上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。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)——连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0079] 值得注意的是,上述一种系统日志价值评价装置的实施例中,所包括的各个单元和模块只是按照功能逻辑进行划分的,但并不局限于上述的划分,只要能够实现相应的功能即可;另外,各功能单元的具体名称也只是为了便于相互区分,并不用于限制本发明的保护范围。

[0080] 注意,上述仅为本发明的较佳实施例及所运用技术原理。本领域技术人员会理解,本发明不限于这里所述的特定实施例,对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本发明的保护范围。因此,虽然通过以上实施例对本发明进行了较为详细的说明,但是本发明不仅仅限于以上实施例,在不脱离本发明构思的情况下,还可以包括更多其他等效实施例,而本发明的范围由所附的权利要求范围决定。

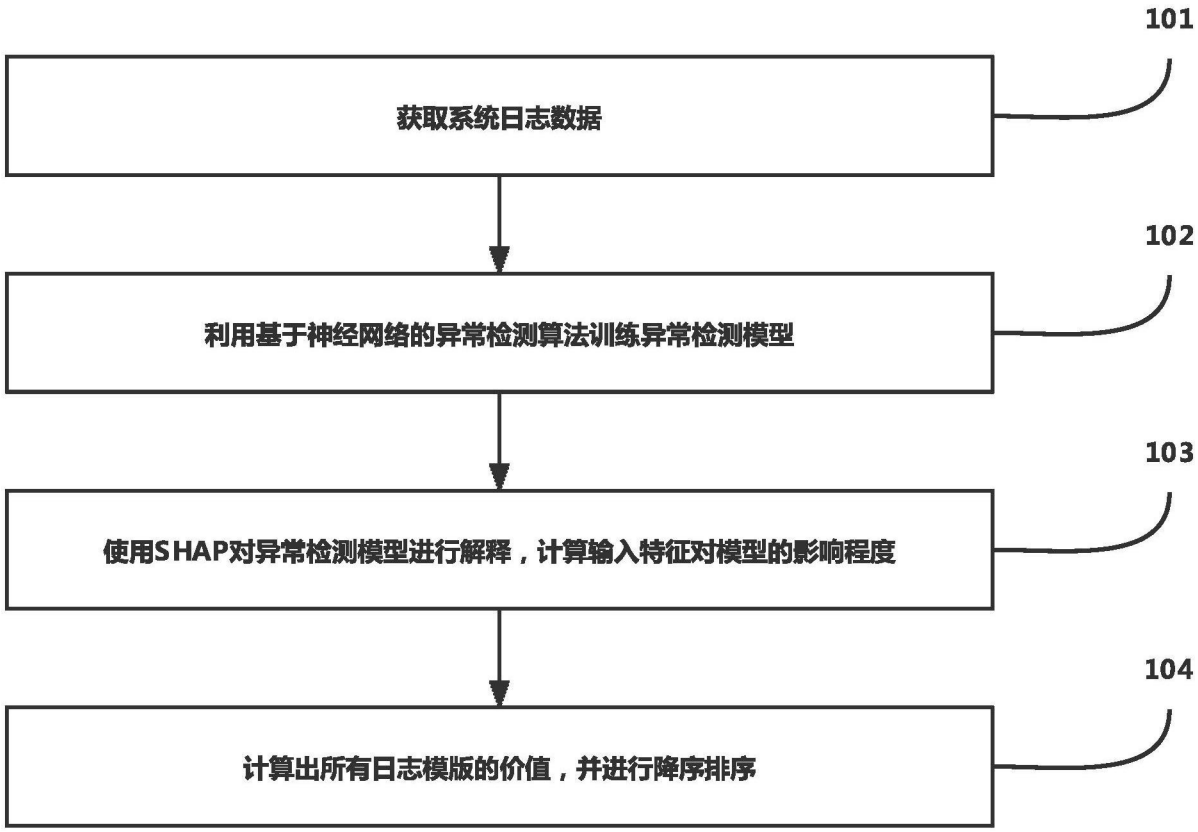

101
获取系统日志数据
102
利用基于神经网络的异常检测算法训练异常检测模型
103
使用SHAP对异常检测模型进行解释，计算输入特征对模型的影响程度
104
计算出所有日志模版的价值，并进行降序排序

图1

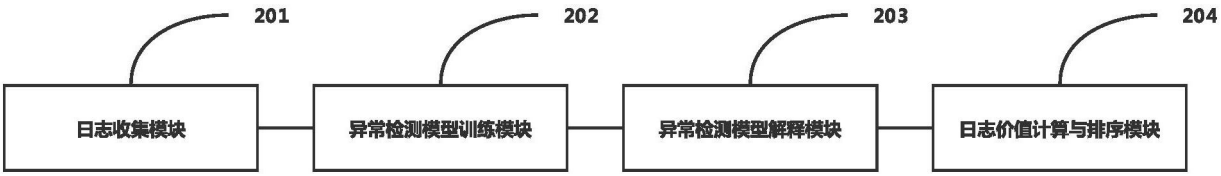

201
日志收集模块
202
异常检测模型训练模块
203
异常检测模型解释模块
204
日志价值计算与排序模块

图2

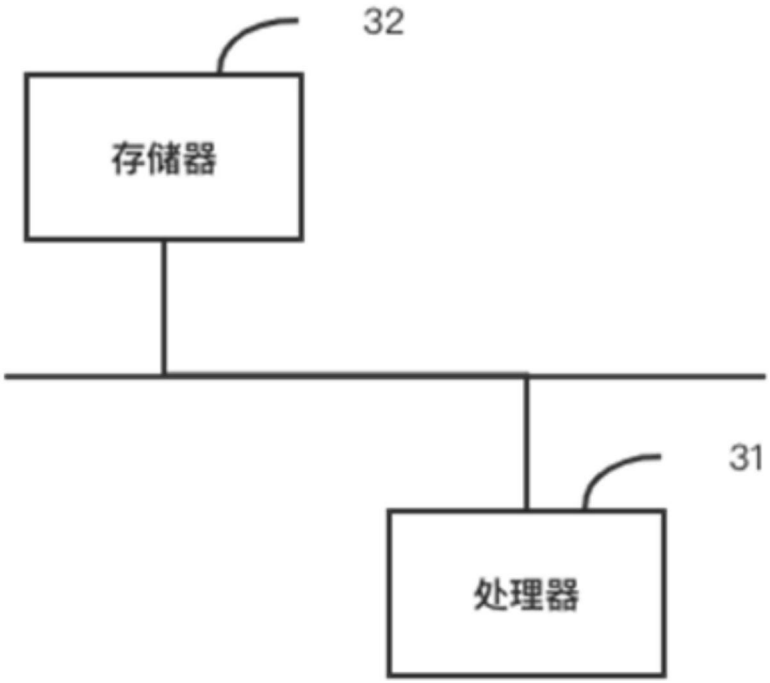


图3