(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110210512 A
(43)申请公布日 2019.09.06

(21)申请号 201910320115.1

(22)申请日 2019.04.19

(71)申请人 北京亿阳信通科技有限公司
地址 100093 北京市海淀区杏石口路99号1幢20302

(72)发明人 丁健

(74)专利代理机构 北京辰权知识产权代理有限公司 11619
代理人 刘广达

(51)Int.Cl.
G06K 9/62(2006.01)
G06N 20/00(2019.01)

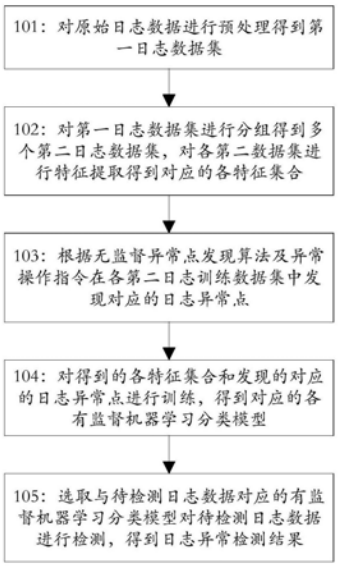
权利要求书3页 说明书8页 附图1页

(54)发明名称

一种自动化日志异常检测方法及系统

(57)摘要

本发明公开一种自动化日志异常检测方法及系统，属于数据处理领域。所述包括：对原始日志数据进行预处理得到第一日志数据集；对第一日志数据集分组得到多个第二日志数据集，对各第二数据集进行特征提取得到对应的各特征集合；根据无监督异常点发现算法及异常操作指令在各第二日志数据集中发现对应的日志异常点；对各特征集合和对应的日志异常点进行训练，得到对应的各有监督机器学习分类模型；选取与待检测日志数据对应的有监督机器学习分类模型对待检测日志数据进行检测，得到日志异常检测结果。本发明中，克服了现有异常检测方法中判别准确性和泛化能力较低、对训练样本中未出现的故障无法预警以及需要耗费极大的时间成本和人工成本的缺陷。



CN 110210512 A

1.一种自动化日志异常检测方法,其特征在于,包括:

步骤S1:对原始日志数据进行预处理得到第一日志数据集;

步骤S2:对所述第一日志数据集进行分组得到多个第二日志数据集,对各第二数据集进行特征提取得到对应的各特征集合;

步骤S3:根据无监督异常点发现算法及异常操作指令在各第二日志数据集中发现对应的日志异常点;

步骤S4:对所述各特征集合和对应的日志异常点进行训练,得到对应的各有监督机器学习分类模型;

步骤S5:选取与待检测日志数据对应的有监督机器学习分类模型对所述待检测日志数据进行检测,得到日志异常检测结果。

2.根据权利要求1所述的方法,其特征在于,所述步骤S1,具体包括:

步骤S1－1:对原始日志数据进行清洗,并保留日志正文和网元类型;

步骤S1－2:对清洗后的原始日志数据进行去参数化及合并处理得到参数泛化日志正文列表;

步骤S1－3:对所述参数泛化日志正文列表中的日志正文分组,并根据各组中各日志正文的长度确定各日志正文的日志模式;

步骤S1－4:根据相同日志模型的日志正文之间的编辑距离确定各日志模式的模板,并根据所述模板确定各日志模式的类型号,得到含有所述网元类型、日志模式类型号的第一日志数据集。

3.根据权利要求2所述的方法,其特征在于,所述步骤S2,具体包括:

步骤S2－1:根据所述网元类型对所述第一日志数据集进行分组,得到各网元类型的第二日志数据集;

步骤S2－2:分别统计在第一预设时间段内各第二日志数据集中每个日志模式的出现次数,得到与所述各第二日志数据集对应的由日志模式类型号和出现次数构成的各特征集合。

4.根据权利要求3所述的方法,其特征在于,所述步骤S3,具体包括:

步骤S3－1:根据无监督异常点发现算法,分别对各第二日志数据集进行训练得到对应的各日志异常点集合;

步骤S3－2:根据历史操作指令集,并结合领域知识构建异常操作指令集;

步骤S3－3:根据所述异常操作指令集,分别对所述各日志异常点集合进行验证,得到各日志异常点集合中各日志异常点的类型。

5.根据权利要求4所述的方法,其特征在于,所述步骤S4,具体包括:

步骤S4－1:根据所述各特征集合和对应的各日志异常点集合构建对应的各异常点训练集;

步骤S4－2:根据随机森林算法,分别对所述各异常点训练集进行训练得到对应的各有监督机器学习分类模型。

6.一种自动化日志异常检测系统,其特征在于,包括:

预处理模块,用于对原始日志数据进行预处理得到第一日志数据集;

分组模块,用于对所述预处理模块得到的第一日志数据集进行分组得到多个第二日志

数据集；

提取模块，用于对所述分组模块得到的各第二数据集进行特征提取得到对应的各特征集合；

异常点发现模块，用于根据无监督异常点发现算法及异常操作指令在所述分组模块得到的各第二日志数据集中发现对应的日志异常点；

训练模块，用于对所述提取模块得到的各特征集合和异常点发现模块发现的对应的日志异常点进行训练，得到对应的各有监督机器学习分类模型；

异常点检测模块，选取与待检测日志数据对应的有监督机器学习分类模型对所述待检测日志数据进行检测，得到日志异常检测结果。

7.根据权利要求6所述的系统，其特征在于，所述预处理模块，包括：清洗子模块、参数化子模块、分组子模块、第一确定子模块和第二确定子模块；

所述清洗子模块，用于对原始日志数据进行清洗，并保留日志正文和网元类型；

所述参数化子模块，用于对所述清洗子模块清洗后的原始日志数据进行去参数化及合并处理得到参数泛化日志正文列表；

所述分组子模块，用于对所述参数化子模块得到的参数泛化日志正文列表中的日志正文分组；

所述第一确定子模块，用于根据所述分组子模块得到的各组中各日志正文的长度确定各日志正文的日志模式；

所述第二确定子模块，用于根据相同日志模型的日志正文之间的编辑距离确定各日志模式的模板，并根据所述模板确定各日志模式的类型号，得到含有所述网元类型、日志模式类型号的第一日志数据集。

8.根据权利要求7所述的系统，其特征在于，

所述分组模块，具体用于：根据所述所述清洗子模块保留的网元类型对所述预处理模块得到的第一日志数据集进行分组，得到各网元类型的第二日志数据集；

所述提取模块，具体用于：分别统计在第一预设时间段内所述分组模块得到的各第二日志数据集中每个日志模式的出现次数，得到与所述各第二日志数据集对应的由日志模式类型号和出现次数构成的各特征集合。

9.根据权利要求8所述的系统，其特征在于，异常点发现模块，具体包括：第一训练子模块、第一构建子模块和验证子模块；

所述第一训练子模块，用于根据无监督异常点发现算法，分别对所述分组模块得到的各第二日志数据集进行训练得到对应的各日志异常点集合；

所述第一构建子模块，用于根据历史操作指令集，并结合领域知识构建异常操作指令集；

所述验证子模块，用于根据所述第一构建子模块构建的异常操作指令集，分别对所述第一训练子模块得到的各日志异常点集合进行验证，得到各日志异常点集合中各日志异常点的类型。

10.根据权利要求9所述的系统，其特征在于，所述训练模块，包括：第二构建子模块和第二训练子模块；

所述第二构建子模块，用于根据所述提取模块得到的各特征集合和所述异常点发现子

模块得到的各日志异常点集合构建对应的各异常点训练集；

所述第二训练子模块，用于根据随机森林算法，分别对所述第二构建子模块构建的异常点训练集进行训练得到对应的各有监督机器学习分类模型。

一种自动化日志异常检测方法及系统

技术领域

[0001] 本发明涉及数据处理领域，尤其涉及一种自动化日志异常检测方法及系统。

背景技术

[0002] 随着技术的快速发展，移动通信系统变得越来越复杂，系统的运营与维护由于需要大量的时间成本、人力成本，已逐渐成为各大移动通信网络运营商的主要支出。因此，实现电信网络设备的自动化异常检测与故障预警，是运营商实现利益最大化的重要途径，并已成为近些年移动通信领域中研究的热点。

[0003] 目前的电信网络设备中，通常存在较为完善的日志记录模块，用于记录诊断日志、操作日志、系统日志等，由于电信网络复杂度的不断提高，目前这些日志数据呈现出以下特点：(1)数据量较大，某运营商的中等省份网络数据产生速率能达到每天9亿条，占据200GByte空间；(2)结构复杂，日志数据设备厂家来源众多，没有标准日志格式模板；(3)正负样本不均，网络告警时期的数据样本占总样本比例低；(4)故障类型多样，单种故障数据样本少，且存在样本中未出现的故障。

[0004] 由于日志数据是电信网络安全状态重要的信息来源，因此其对网络故障预警具有重要意义。当前利用日志数据进行故障预警的方法有很多，主要包括：统计学方法、基于机器学习的方法以及基于专家知识的异常检测方法。其中，统计学方法适用于正常行为统计模型，通过对测试数据进行测试，给出异常分数，如果异常分数高于一个阈值，则认为是异常点；该方法在设置恰当的阈值以及调整好参数的前提下，可以提供较准确的预测。基于机器学习的方法，主要包括分类算法和聚类算法；其中，分类算法是一种有监督的机器学习算法，其必要前提是训练集包含的分类数据所属类别是已知的；而聚类算法是一种无监督的机器学习算法，通常是基于距离对样本数据进行聚类，识别出异常点，但此种方法存在对训练样本中未出现的故障无法预警的缺陷。基于专家知识的异常检测，又称为专家系统，专家系统是以规则为基础，利用预定义的规则对测试数据进行匹配，并可以不断获取知识，进入一个更高的置信区域，根据分数阈值，判定异常行为。同时，基于机器学习与专家知识相结合的方法在计算机数据管理技术领域也同样有所应用，其是基于系统的源代码分析，对程序的运行日志提取与性能相关特征向量，并结合机器学习算法和专家知识，有效检测和诊断程序的常见性能异常。

[0005] 目前，上述方法均有实际的应用，并且存在相关的专利申请；其中，基于统计学原理来进行故障预警的技术方案可参见申请号为CN201410191589.8、CN201510765610.5和CN201611213764.4的专利；基于机器学习进行故障预警的技术方案可参见申请号为CN201610125901.2和CN201611232408.7的专利；基于知识的异常检测技术方案可参见申请号为201510180528.6的专利；基于机器学习方法与专家知识相结合的技术方案可参见申请号为CN201610312729.1的专利。

[0006] 然而，上述方法并不完善，其中，基于统计学的异常检测方法，虽然在设置恰当的阈值以及调整好参数的前提下，可以提供较准确的预测，但是阈值以及参数的调试是非常

困难的，模型训练需要耗费很长时间，此外在训练模型时，每个变量都被假设是满足统计分布的，大多数训练方案也依赖于一个假设过程，而这是不现实的。基于机器学习的异常检测方法，在其日志数据正负样本不均，单种故障数据样本少等情况下，判别准确性和泛化的能力较低，且对训练样本中未出现的故障无法预警。基于知识的异常检测方法，其高质量规则库的建立将会耗费极大的时间成本和人工成本，而且此方法难以检测罕见的、未知的异常。基于机器学习与专家系统相结合的方法，虽然使用专家知识对异常类别进行标注，提高了判别准确度，但是同样需要耗费极大的时间成本和人工成本。

[0007] 可见，目前仍没有一个完善的方法检测日志数据中的异常，进而进行故障的预警。

发明内容

[0008] 为解决现有技术的不足，本发明提供一种自动化日志异常检测方法及系统。

[0009] 一方面，本发明提供一种自动化日志异常检测方法，包括：

[0010] 步骤S1：对原始日志数据进行预处理得到第一日志数据集；

[0011] 步骤S2：对所述第一日志数据集进行分组得到多个第二日志数据集，对所述第二数据集进行特征提取得到对应的各特征集合；

[0012] 步骤S3：根据无监督异常点发现算法及异常操作指令在各第二日志数据集中发现日志异常点；

[0013] 步骤S4：对所述各特征集合和对应的日志异常点进行训练，得到对应的各有监督机器学习分类模型；

[0014] 步骤S5：选取与待检测日志数据对应的有监督机器学习分类模型对所述待检测日志数据进行检测，得到日志异常检测结果。

[0015] 可选地，所述步骤S1，具体包括：

[0016] 步骤S1－1：对原始日志数据进行清洗，并保留日志正文和网元类型；

[0017] 步骤S1－2：对清洗后的原始日志数据进行去参数化及合并处理得到参数泛化日志正文列表；

[0018] 步骤S1－3：对所述参数泛化日志正文列表中的日志正文分组，并根据各组中各日志正文的长度确定各日志正文的日志模式；

[0019] 步骤S1－4：根据相同日志模型的日志正文之间的编辑距离确定各日志模式的模板，并根据所述模板确定各日志模式的类型号，得到含有所述网元类型、日志模式类型号的第一日志数据集。

[0020] 可选地，所述步骤S2，具体包括：

[0021] 步骤S2－1：根据所述网元类型对所述第一日志数据集进行分组，得到各网元类型的第二日志数据集；

[0022] 步骤S2－2：分别统计在第一预设时间段内各第二日志数据集中每个日志模式的出现次数，得到与所述各第二日志数据集对应的由日志模式类型号和出现次数构成的各特征集合。

[0023] 可选地，所述步骤S3，具体包括：

[0024] 步骤S3－1：根据无监督异常点发现算法，分别对各第二日志数据集进行训练得到对应的各日志异常点集合；

[0025] 步骤S3－2:根据历史操作指令集,并结合领域知识构建异常操作指令集;

[0026] 步骤S3－3:根据所述异常操作指令集,分别对所述各日志异常点集合进行验证,得到所述各日志异常点集合中各日志异常点的类型。

[0027] 可选地,所述步骤S4,具体包括:

[0028] 步骤S4－1:根据所述各特征集合和对应的所述各日志异常点构建对应的各异常点训练集;

[0029] 步骤S4－2:根据随机森林算法,分别对所述各异常点训练集进行训练得到对应的各有监督机器学习分类模型。

[0030] 另一方面,本发明提供一种自动化日志异常检测系统,包括:

[0031] 预处理模块,用于对原始日志数据进行预处理得到第一日志数据集;

[0032] 分组模块,用于对所述预处理模块得到的第一日志数据集进行分组得到多个第二日志数据集;

[0033] 提取模块,用于对所述分组模块得到的第二数据集进行特征提取得到对应的各特征集合;

[0034] 异常点发现模块,用于根据无监督异常点发现算法及异常操作指令在所述分组模块得到的各第二日志训练数据集中发现对应的日志异常点;

[0035] 训练模块,用于对所述提取模块得到的各特征集合和所述异常点发现模块发现的对应的日志异常点进行训练,得到对应的各有监督机器学习分类模型;

[0036] 异常点检测模块,用于选取与待检测日志数据对应的有监督机器学习分类模型对所述待检测日志数据进行检测,得到日志异常检测结果。

[0037] 可选地,所述预处理模块,具体包括:清洗子模块、参数化子模块、分组子模块、第一确定子模块和第二确定子模块;

[0038] 所述清洗子模块,用于对原始日志数据进行清洗,并保留日志正文和网元类型;

[0039] 所述参数化子模块,用于对所述清洗子模块清洗后的原始日志数据进行去参数化及合并处理得到参数泛化日志正文列表;

[0040] 所述分组子模块,用于对所述参数化子模块得到的参数泛化日志正文列表中的日志正文分组;

[0041] 所述第一确定子模块,用于根据所述分组子模块得到的各组中各日志正文的长度确定各日志正文的日志模式;

[0042] 所述第二确定子模块,用于根据相同日志模型的日志正文之间的编辑距离确定各日志模式的模板,并根据所述模板确定各日志模式的类型号,得到含有所述网元类型、日志模式类型号的第一日志数据集。

[0043] 可选地,所述分组模块,具体用于:根据所述所述清洗子模块保留的网元类型对所述预处理模块得到的第一日志数据集进行分组,得到各网元类型的第二日志数据集;

[0044] 可选地,所述提取模块,具体用于:分别统计在第一预设时间段内所述分组模块得到的各第二日志数据集中每个日志模式的出现次数,得到与所述各第二日志数据集对应的由日志模式类型号和出现次数构成的各特征集合。

[0045] 可选地,异常点发现模块,具体包括:第一训练子模块、第一构建子模块和验证子模块;

[0046] 所述第一训练子模块，用于根据无监督异常点发现算法，分别对所述分组模块得到的各第二日志数据集进行训练得到对应的各日志异常点集合；

[0047] 所述第一构建子模块，用于根据历史操作指令集，并结合领域知识构建异常操作指令集；

[0048] 所述验证子模块，用于根据所述第一构建子模块构建的异常操作指令集，分别对所述第一训练子模块得到的各日志异常点集合进行验证，得到各日志异常点集合中各日志异常点的类型。

[0049] 可选地，所述训练模块，具体包括：第二构建子模块和第二训练子模块；

[0050] 所述第二构建子模块，用于根据所述提取模块得到的各特征集合和所述异常点发现子模块得到的各日志异常点构建对应的各异常点训练集；

[0051] 所述第二训练子模块，用于根据随机森林算法，分别对所述第二构建子模块构建的各异常点训练集进行训练得到对应的各有监督机器学习分类模型。

[0052] 本发明的优点在于：

[0053] 本申请通过在原始日志数据中提取特征，使用无监督异常点发现算法并结合异常操作指令发现异常点，进而基于异常点训练出有监督机器学习分类模型，通过有监督机器学习分类模型实现待检测日志数据中异常点的自动化检测，进而进行故障预警；不仅克服了基于机器学习的异常检测方法中判别准确性和泛化能力较低、对训练样本中未出现的故障无法预警的缺陷，也克服了基于知识的异常检测方法中需要耗费极大的时间成本和人工成本的缺陷。

附图说明

[0054] 通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

[0055] 附图1为本发明提供的一种自动化日志异常检测方法流程图；

[0056] 附图2为本发明提供的一种自动化日志异常检测系统模块组成框图。

具体实施方式

[0057] 下面将参照附图更详细地描述本公开的示例性实施方式。虽然附图中显示了本公开的示例性实施方式，然而应当理解，可以以各种形式实现本公开而不应被这里阐述的实施方式所限制。相反，提供这些实施方式是为了能够更透彻地理解本公开，并且能够将本公开的范围完整的传达给本领域的技术人员。

[0058] 实施例一

[0059] 根据本发明的实施方式，提供一种自动化日志异常检测方法，如图1所示，包括：

[0060] 步骤101：对原始日志数据进行预处理得到第一日志数据集；

[0061] 在本实施例中，步骤101，具体包括：

[0062] 步骤101－1：对原始日志数据进行清洗，并保留日志正文和网元类型；

[0063] 具体地，对原始日志数据进行清洗，去除冗余字符，并保留网元类型、日志时间、日志类型、日志正文等关键信息。

[0064] 步骤101－2:对清洗后的原始日志数据进行去参数化及合并处理得到参数泛化日志正文列表;

[0065] 具体地,采用正则表达式匹配的方式,将清洗后的原始日志数据中各日志正文含有的数值参数替换为占位符,实现去参数化处理,并将去参数化处理后具有相同结构的日志文正进行合并,得到参数泛化日志正文列表;

[0066] 例如,在本实施例中,将日志正文中含有的日期、IP地址、电话号码、URL等信息替换为占位符。

[0067] 步骤101－3:对参数泛化日志正文列表中的日志正文分组,并根据各组中各日志正文的长度确定各日志正文的日志模式;

[0068] 具体地,根据日志正文的文本长度对参数泛化日志正文列表中的日志正文进行分组;并计算各组中任意两个文本长度相同的日志正文之间的编辑距离,将编辑距离小于预设阈值的日志正文归为同一种日志模式。

[0069] 步骤101－4:根据相同日志模型的日志正文之间的编辑距离确定各日志模式的模板,并根据确定的模板确定各日志模式的类型号,得到含有网元类型、日志模式类型号的第一日志数据集。

[0070] 具体地,分别在每种日志模式中选取与该日志模式中其他日志正文的平均编辑距离最小的日志正文作为该日志模式的模板,并将作为模板的日志正文的哈希值作为该日志模式的类型号,得到含有日志时间、日志类型、网元类型、日志模式类型号的第一日志数据集。

[0071] 步骤102:对第一日志数据集进行分组得到多个第二日志数据集,对各第二数据集进行特征提取得到对应的各特征集合;

[0072] 在本实施例中,步骤102,具体包括:

[0073] 步骤102－1:根据网元类型对第一日志数据集进行分组,得到各网元类型的第二日志数据集;

[0074] 本发明中,由于不同网元类型对应的日志数据的差别较大,故根据网元类型对第一日志数据集进行分组,进而进行后续操作以得到各网元类型的日志异常点检测模型,即有监督机器学习分类模型。

[0075] 步骤102－2:分别统计在第一预设时间段内各第二日志数据集中每个日志模式的出现次数,得到与各第二日志数据集对应的由日志模式类型号和出现次数构成的各特征集合。

[0076] 具体地,任意选取各网元类型中的一种网元类型,统计在第一预设时间段内选取的网元类型的第二日志数据集中的每个日志模式的出现次数,得到与选取的网元类型的第二日志数据集对应的由日志模式类型号和出现次数构成的特征集合;重复上述操作,直至得到所有第二日志数据集对应的各特征集合。

[0077] 步骤103:根据无监督异常点发现算法及异常操作指令在各第二日志训练数据集中发现对应的日志异常点;

[0078] 在本实施例中,步骤103,具体包括:

[0079] 步骤103－1:根据无监督异常点发现算法,分别对各第二日志数据集进行训练得到对应的各日志异常点集合;

[0080] 本实施例中，无监督异常点发现算法，例如为K－means等聚类算法。

[0081] 步骤103－2：根据历史操作指令集，并结合领域知识构建异常操作指令集；

[0082] 具体地，根据历史运维人员所使用的操作指令集，并结合领域知识构建异常操作指令集。

[0083] 步骤103－3：根据构建的异常操作指令集，分别对各日志异常点集合进行验证，得到各日志异常点集合中各日志异常点的类型。

[0084] 具体地，依次判断各日志异常点集合中各日志异常点在其产生后的第二预设时间段内，是否有运维人员对该日志异常点执行了异常操作指令集中的指令，是则判定该日志异常点的类型为有效日志异常点；否则，则判定该日志异常点的类型为无效日志异常点。

[0085] 步骤104：对得到的各特征集合和发现的对应的日志异常点进行训练，得到对应的各有监督机器学习分类模型；

[0086] 根据本发明的实施方式，步骤104，具体包括：

[0087] 步骤104－1：根据各特征集合和对应的各日志异常点集合构建对应的各异常点训练集；

[0088] 其中，异常点训练集中包括：各日志异常点的时间戳(日志时间)、日志模式类型号、日志模式类型号出现次数、异常点类型等信息。

[0089] 步骤104－2：根据随机森林算法，分别对各异常点训练集进行训练得到对应的各有监督机器学习分类模型。

[0090] 步骤105：选取与待检测日志数据对应的有监督机器学习分类模型对待检测日志数据进行检测，得到日志异常检测结果。

[0091] 具体地，根据待检测日志数据的网元类型，选取对应的有监督机器学习分类模型对待检测日志数据进行检测，得到日志异常检测结果。

[0092] 实施例二

[0093] 根据本发明的实施方式，提供一种自动化日志异常检测系统，如图2所示，包括：

[0094] 预处理模块201，用于对原始日志数据进行预处理得到第一日志数据集；

[0095] 分组模块202，用于对预处理模块201得到的第一日志数据集进行分组得到多个第二日志数据集；

[0096] 提取模块203，用于对分组模块202得到的各第二数据集进行特征提取得到对应的各特征集合；

[0097] 异常点发现模块204，用于根据无监督异常点发现算法及异常操作指令在分组模块202得到的各第二日志训练数据集中发现对应的日志异常点；

[0098] 训练模块205，用于对提取模块203得到的各特征集合和异常点发现模块204发现的对应的日志异常点进行训练，得到对应的各有监督机器学习分类模型；

[0099] 异常点检测模块206，选取与待检测日志数据对应的训练模块205得到的有监督机器学习分类模型对待检测日志数据进行检测，得到日志异常检测结果。

[0100] 根据本发明的实施方式，预处理模块201，具体包括：清洗子模块、参数化子模块、分组子模块、第一确定子模块和第二确定子模块，其中：

[0101] 清洗子模块，用于对原始日志数据进行清洗，并保留日志正文和网元类型；

[0102] 参数化子模块，用于对清洗子模块清洗后的原始日志数据进行去参数化及合并处

理得到参数泛化日志正文列表；

[0103] 分组子模块，用于将参数化子模块得到的参数泛化日志正文列表中的日志正文进行分组；

[0104] 第一确定子模块，用于根据分组子模块得到的各组中各日志正文的长度确定各日志正文的日志模式；

[0105] 第二确定子模块，用于根据相同日志模型的日志正文之间的编辑距离确定各日志模式的模板，并根据所述模板确定各日志模式的类型号，得到含有网元类型、日志模式类型号的第一日志数据集。

[0106] 进一步地，在本实施例中，参数化子模块，具体用于：采用正则表达式匹配的方式，将清洗子模块清洗后的原始日志数据中各日志正文含有的数值参数替换为占位符，实现去参数化处理，并将去参数化处理后具有相同结构的日志文正进行合并，得到参数泛化日志正文列表。

[0107] 根据本发明的实施方式，分组模块202，具体用于：根据清洗子模块保留的网元类型对预处理模块201得到的第一日志数据集进行分组，得到各网元类型的第二日志数据集；

[0108] 提取模块203，具体用于：分别统计在第一预设时间段内分组模块202得到的各第二日志数据集中每个日志模式的出现次数，得到与各第二日志数据集对应的由日志模式类型号和出现次数构成的各特征集合。

[0109] 根据本发明的实施方式，异常点发现模块204，具体包括：第一训练子模块、第一构建子模块和验证子模块，其中：

[0110] 第一训练子模块，用于根据无监督异常点发现算法，分别对分组模块202得到的各第二日志数据集进行训练得到对应的各日志异常点集合；

[0111] 第一构建子模块，用于根据历史操作指令集，并结合领域知识构建异常操作指令集；

[0112] 验证子模块，用于根据第一构建子模块构建的异常操作指令集，分别对第一训练子模块得到的各日志异常点集合进行验证，得到各日志异常点集合中各日志异常点的类型。

[0113] 根据本发明的实施方式，训练模块205，具体包括：第二构建子模块和第二训练子模块，其中：

[0114] 第二构建子模块，用于根据提取模块203得到的各特征集合和异常点发现子模块得到的各日志异常点集合构建对应的各异常点训练集；

[0115] 第二训练子模块，用于根据随机森林算法，分别对第二构建子模块构建的各异常点训练集进行训练得到对应的各有监督机器学习分类模型。

[0116] 根据本发明的实施方式，异常点检测模块206，具体用于：根据待检测日志数据的网元类型，选取对应的有监督机器学习分类模型对待检测日志数据进行检测，得到日志异常检测结果

[0117] 本申请通过在原始日志数据中提取特征，使用无监督异常点发现算法并结合异常操作指令发现异常点，进而基于异常点训练出有监督机器学习分类模型，通过有监督机器学习分类模型实现待检测日志数据中异常点的自动化检测，进而进行故障预警；不仅克服了基于机器学习的异常检测方法中判别准确性和泛化能力较低、对训练样本中未出现的故

障无法预警的缺陷，也克服了基于知识的异常检测方法中需要耗费极大的时间成本和人工成本的缺陷。

[0118] 以上所述，仅为本发明较佳的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到的变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应以所述权利要求的保护范围为准。

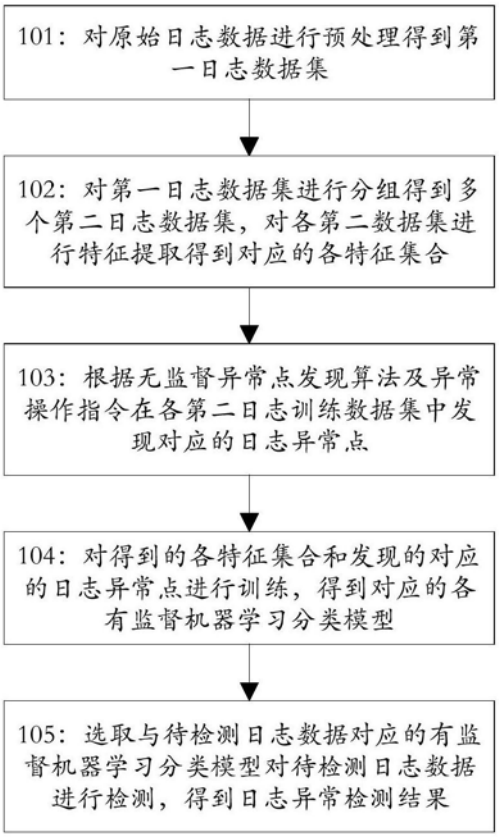

101：对原始日志数据进行预处理得到第一日志数据集
102：对第一日志数据集进行分组得到多个第二日志数据集，对各第二数据集进行特征提取得到对应的各特征集合
103：根据无监督异常点发现算法及异常操作指令在各第二日志训练数据集中发现对应的日志异常点
104：对得到的各特征集合和发现的对应的日志异常点进行训练，得到对应的各有监督机器学习分类模型
105：选取与待检测日志数据对应的有监督机器学习分类模型对待检测日志数据进行检测，得到日志异常检测结果

图1

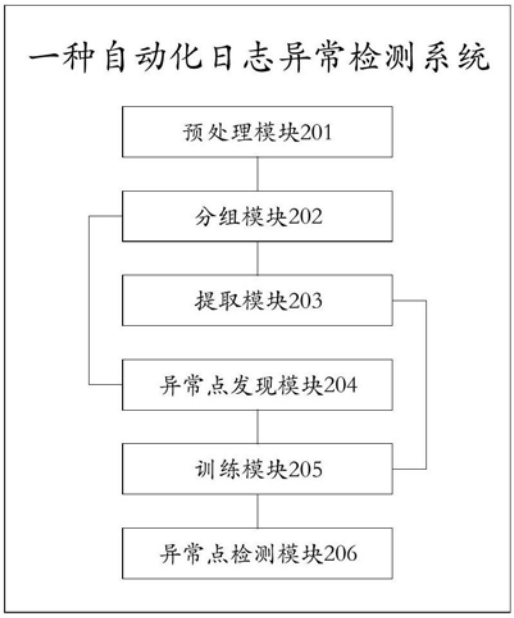

一种自动化日志异常检测系统
预处理模块201
分组模块202
提取模块203
异常点发现模块204
训练模块205
异常点检测模块206

图2