(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 112966222 A
(43)申请公布日 2021.06.15

(21)申请号 202110261184.7

(22)申请日 2021.03.10

(71)申请人 中国民航信息网络股份有限公司
地址 100085 北京市顺义区后沙峪镇裕民大街7号

(72)发明人 郭东丹 刘晓辉 周子站 周凯洋 李婷

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 柳欣

(51)Int.Cl.
G06F 17/18(2006.01)
G06K 9/62(2006.01)

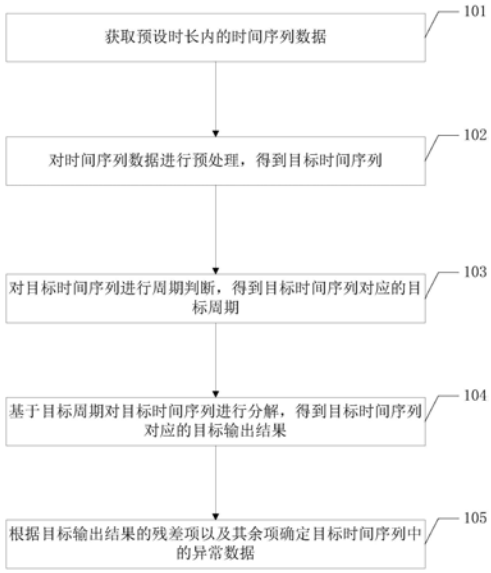
权利要求书4页 说明书13页 附图5页

(54)发明名称

一种时间序列异常数据检测方法及相关设备

(57)摘要

本申请提供了一种时间序列异常数据检测方法及相关设备，可以不需要大量的特征工程和人工标注，也无需大量内存用于存储向量描述，减少工作量及内存成本，同时降低人力成本。该方法包括：获取预设时长内的时间序列数据；对所述时间序列数据进行预处理，得到目标时间序列；对所述目标时间序列进行周期判断，得到所述目标时间序列对应的目标周期；基于所述目标周期对所述目标时间序列进行分解，得到所述目标时间序列对应的目标输出结果，所述目标输出结果包括残差项以及其余项；根据所述目标输出结果的残差项以及其余项确定所述目标时间序列中的异常数据。



CN 112966222 A

1.一种时间序列异常数据检测方法,其特征在于,包括:

获取预设时长内的时间序列数据;

对所述时间序列数据进行预处理,得到目标时间序列;

对所述目标时间序列进行周期判断,得到所述目标时间序列对应的目标周期;

基于所述目标周期对所述目标时间序列进行分解,得到所述目标时间序列对应的目标输出结果,所述目标输出结果包括残差项以及其余项;

根据所述目标输出结果的残差项以及其余项确定所述目标时间序列中的异常数据。

2.根据权利要求1所述的方法,其特征在于,所述对所述目标时间序列进行周期判断,得到所述目标时间序列对应的目标周期包括:

通过如下公式计算所述目标时间序列对应的目标周期:

$$T=\arg\max_{l} ACF(l);$$

其中,T为所述目标周期,ACF(l)为所述目标时间序列对应的不同的时间延迟l的自相关系数/序列相关函数,通过如下公式计算ACF(l):

$$ACF(l)=\frac{\sum_{i=1}^{n} Y_i Y_{i+l}}{n}, l=1,2,\cdots,\frac{n}{2};$$

其中,n为所述目标时间序列中数据点的数量,Y_i为所述目标时间序列中的第i个序列。

3.根据权利要求1所述的方法,其特征在于,所述其余项包括季节项以及趋势项,所述根据所述目标输出结果的残差项以及其余项确定所述目标时间序列中的异常数据包括:

根据所述目标周期计算第一目标数据点的平均值,所述第一目标数据点为所述目标输出结果的残差项中任意一个数据;

计算第一数据点与第二数据点的目标和值,所述第一数据点为所述目标输出结果的季节项与所述第一目标数据点对应的数据点,所述第二数据点为所述目标输出结果的趋势项中与所述第一目标数据点对应的数据点;

根据所述第一目标数据点的平均值以及所述目标和值确定第一基线点、第一上界线点以及第一下界线点,其中,所述第一动态基线点为第一动态基线中与所述第一目标数据点对应的基线点,所述第一上界线点为第一上界线中与所述第一目标数据点对应的上界线点,所述第一下界线点为第一下界线中与所述第一目标数据点对应的下界线点;

将所述目标时间序列中不处于所述第一上界线与所述第一下界线之间的数据点确定为所述目标时间序列的异常数据。

4.根据权利要求3所述的方法,其特征在于,所述根据所述目标周期计算第一目标数据点的平均值包括:

通过如下公式计算所述第一目标数据点的平均值:

$$M_i=\begin{cases}\dfrac{1}{T}\sum_{i=1}^{T}R_i, i=T\\ M_{i+1}+\dfrac{(R_i-R_{i-T})}{T}, T<i\le n\end{cases};$$

其中，M_i为所述第一目标数据点的平均值，R_i为所述第一目标数据点，T为所述目标周期；

所述根据所述第一目标数据点的平均值以及所述目标和值确定第一基线点、第一上界线点以及第一下界线点包括：

通过如下公式计算所述第一基线点：

$B_i=M_i+A_i$；

其中，B_i为所述第一基线点，A_i为所述目标和值，$A_i=T_i+S_i$，T_i为所述第二数据点，S_i为所述第一数据点；

通过如下公式计算所述第一上界线点：

$$B_i^{upper}=A_i+M_i+3\times\sigma_i;$$

其中，B_i^{upper}为所述第一上界线点，σ_i为所述第一目标数据点对应的标准差，

$$\sigma_i=\sqrt{\frac{1}{t}\sum_{j=i-T+1}^{i}(R_j-M_j)^2};$$

通过如下公式计算所述第一下界线点：

$$B_i^{lower}=A_i+M_i-3\times\sigma_i;$$

其中，B_i^{lower}为所述第一下界线点。

5.根据权利要求1所述的方法，其特征在于，所述根据所述目标输出结果的残差项以及其余项确定所述目标时间序列中的异常数据包括：

确定第二目标数据点的第一四分位数、中位数以及第三四分位数，所述第二目标数据点为所述目标输出结果的残差项中的任意一个数据点；

确定所述其余项中与所述第一目标数据点对应的第三数据点的值；

根据所述中位数以及所述第三数据点确定第二基线点，所述第二基线点为第二动态基线中与所述目标数据点对应的基线点；

根据所述第一四分位数、所述第三四分位数以及所述第三数据点计算第二上界线点以及第二下界线点，所述第二上界线点为第二上界线中与所述目标数据点对应的上界线点，所述第二下界线点为第二下界线中与所述目标数据点对应的下界线点；

将所述目标时间序列中不处于所述第二上界线与所述第二下界线之间的数据点确定为所述目标时间序列的异常数据。

6.根据权利要求5所述的方法，其特征在于，所述根据所述中位数以及所述第三数据点

确定第二基线点包括：

通过如下公式计算所述第二基线点：

$B_{i'}=Q_{2(i')}+A_{i'}$；

其中，$B_{i'}$为所述第二基线点，$A_{i'}$为所述第三数据点，$Q_{2(i')}$为所述中位数；

所述根据所述第一四分位数、所述第三四分位数以及所述第三数据点计算第二上界线点以及第二下界线点包括：

通过如下公式计算所述第二上界线点：

$$B_{i'}^{upper}=A_{i'}+Q_{3(i')}+k\times(Q_{3(i')}-Q_{1(i')})\ ;$$

其中，$B_{i'}^{upper}$为所述第二上界线点，$Q_{3(i')}$为所述第三四分位数，$Q_{1(i')}$为所述第一四分位数，k为常数；

通过如下公式计算所述第二下界线点：

$$B_{i'}^{lower}=A_{i'}+Q_{1(i')}-k\times(Q_{3(i')}-Q_{1(i')})\ ;$$

其中，$B_{i'}^{lower}$为所述第二下界线点。

7.一种时间序列异常数据检测装置，其特征在于，包括：

获取单元，用于获取预设时长内的时间序列数据；

预处理单元，用于对所述时间序列数据进行预处理，得到目标时间序列；

周期判断单元，用于对所述目标时间序列进行周期判断，得到所述目标时间序列对应的目标周期；

分解单元，用于基于所述目标周期对所述目标时间序列进行分解，得到所述目标时间序列对应的目标输出结果，所述目标输出结果包括残差项以及其余项；

确定单元，用于根据所述目标输出结果的残差项以及其余项确定所述目标时间序列中的异常数据。

8.根据权利要求7所述的装置，其特征在于，所述其余项包括季节项以及趋势项，所述确定单元具体用于：

根据所述目标周期计算第一目标数据点的平均值，所述第一目标数据点为所述目标输出结果的残差项中任意一个数据；

计算第一数据点与第二数据点的目标和值，所述第一数据点为所述目标输出结果的季节项与所述第一目标数据点对应的数据点，所述第二数据点为所述目标输出结果的趋势项中与所述第一目标数据点对应的数据点；

根据所述第一目标数据点的平均值以及所述目标和值确定第一基线点、第一上界线点以及第一下界线点，其中，所述第一动态基线点为第一动态基线中与所述第一目标数据点对应的基线点，所述第一上界线点为第一上界线中与所述第一目标数据点对应的上界线点，所述第一下界线点为第一下界线中与所述第一目标数据点对应的下界线点；

将所述目标时间序列中不处于所述第一上界线与所述第一下界线之间的数据点确定为所述目标时间序列的异常数据。

9.根据权利要求7所述的装置，其特征在于，所述确定单元还具体用于：

确定第二目标数据点的第一四分位数、中位数以及第三四分位数，所述第二目标数据点为所述目标输出结果的残差项中的任意一个数据点；

确定所述其余项中与所述第一目标数据点对应的第三数据点的值；

根据所述中位数以及所述第三数据点确定第二基线点，所述第二基线点为第二动态基线中与所述目标数据点对应的基线点；

根据所述第一四分位数、所述第三四分位数以及所述第三数据点计算第二上界线点以及第二下界线点，所述第二上界线点为第二上界线中与所述目标数据点对应的上界线点，所述第二下界线点为第二下界线中与所述目标数据点对应的下界线点；

将所述目标时间序列中不处于所述第二上界线与所述第二下界线之间的数据点确定为所述目标时间序列的异常数据。

10.一种机器可读介质，其特征在于，包括指令，当所述指令在机器上运行时，使得机器执行上述权利要求1至6中任一项所述的时间序列异常数据检测方法的步骤。

一种时间序列异常数据检测方法及相关设备

技术领域

[0001] 本申请涉及通信领域,尤其涉及一种时间序列异常数据检测方法及相关设备。

背景技术

[0002] 在民航业务信息系统中,时间序列异常检测是保证服务质量的重要手段。随着服务的规模和复杂度不断增加,监测系统的各种关键性能指标(KPI)及时发现异常并处理异常,可以防止由异常未及时处理导致的服务器瘫痪,避免损害业务和影响客户体验的情况发生。在运维管理系统中,由于异常检测需要具有实时性,人工检测的方法会耗费大量的人力,一般设法实现自动异常检测。

[0003] 现有的时间序列异常检测的方法是将时序异常检测作为一个二分类问题,通过两种方法来进行异常检测:一种是传统的机器学习方法,另外一种是基于深度神经网络的方法。

[0004] 传统的机器学习方法需要使用统计分析或者其他方法提取大量的特征,然后将这些特征输入强大的分类器中,比如一类支持向量机(One Class-Support vector machine,OC-SVM),随机森林(Random Forest,RF),支持向量数据描述(Support Vector Data Description,SVDD)等,这类方法的局限性在于计算扩展性差和维度灾难,在高维数据的场景中不适用,且使用该类方法需要存储大量的向量数据描述,需要大量内存。

[0005] 基于深度神经网络的方法虽然不用经过提取大量特征的过程,但是它作为有监督学习方法,在异常检测问题中,绝大多数样本为正常样本,只有少数样本为异常样本,这样容易导致模型的学习效果很差,且需要人工在大量的样本中找到异常样本打上标签用来训练模型,增加工作量以及人力成本。

发明内容

[0006] 本申请提供了一种时间序列异常数据检测方法及相关设备,无需经过大量的特征提取过程,适用于高维度时间序列数据的场景,适配只有少量样本为异常样本的场景,且不需要人工给异常样本打标签。

[0007] 本申请实施例第一方面提供了一种时间序列异常数据检测方法,包括:

[0008] 获取预设时长内的时间序列数据;

[0009] 对所述时间序列数据进行预处理,得到目标时间序列;

[0010] 对所述目标时间序列进行周期判断,得到所述目标时间序列对应的目标周期;

[0011] 基于所述目标周期对所述目标时间序列进行分解,得到所述目标时间序列对应的目标输出结果,所述目标输出结果包括残差项以及其余项;

[0012] 根据所述目标输出结果的残差项以及其余项确定所述目标时间序列中的异常数据。

[0013] 本申请实施例第二方面提供了一种时间序列异常数据检测装置,包括:

[0014] 获取单元,用于获取预设时长内的时间序列数据;

[0015] 预处理单元，用于对所述时间序列数据进行预处理，得到目标时间序列；

[0016] 周期判断单元，用于对所述目标时间序列进行周期判断，得到所述目标时间序列对应的目标周期；

[0017] 分解单元，用于基于所述目标周期对所述目标时间序列进行分解，得到所述目标时间序列对应的目标输出结果，所述目标输出结果包括残差项以及其余项；

[0018] 确定单元，用于根据所述目标输出结果的残差项以及其余项确定所述目标时间序列中的异常数据。

[0019] 本申请第三方面提供了一种计算机装置，其包括至少一个连接的处理器和存储器，其中，所述存储器用于存储程序代码，所述程序代码由所述处理器加载并执行以实现上述第一方面所述的时间序列异常数据检测方法的步骤。

[0020] 本申请实施例第四方面提供了一种机器可读介质，其包括指令，当其在机器上运行时，使得机器执行上述第一方面所述的时间序列异常数据检测方法的步骤。

[0021] 综上所述，可以看出，本申请提供的实施例中，时间序列异常数据检测装置对预设时长内的时间序列数据进行预处理，并对预处理后的时间序列数据进行周期判断得到周期，并基于周期进行分解，得到时间序列的残差项和其余项，之后根据残差项和其余项确定目标时间序列中的异常数据。由此，可以不需要大量的特征工程和人工标注，也无需大量内存用于存储向量描述，减少工作量及内存成本，降低人力成本。

附图说明

[0022] 结合附图并参考以下具体实施方式，本申请各实施例的上述和其他特征、优点及方面将变得更加明显。贯穿附图中，相同或相似的附图标记表示相同或相似的元素。应当理解附图是示意性的，原件和元素不一定按照比例绘制。

[0023] 图1为本申请实施例提供的时间序列异常数据检测方法的一个流程示意图；

[0024] 图2为本申请实施例提供的STL分解效果示意图；

[0025] 图3为本申请实施例提供STL分解算法以及Nsigma算法相结合得到时间序列异常数据的示意图；

[0026] 图4为本申请实施例提供的EMD分解算法以及Nsigma算法相结合得到时间序列异常数据的示意图；

[0027] 图5为本申请实施例提供STL分解算法以及四分位数算法相结合得到时间序列异常数据的示意图；

[0028] 图6为本申请实施例提供的EMD分解算法以及四分位数算法相结合得到时间序列异常数据的示意图；

[0029] 图7为本申请实施例提供的时间序列异常数据检测装置的虚拟结构示意图；

[0030] 图8为本请实施例提供的机器可读介质的结构示意图；

[0031] 图9为本申请实施例提供的服务器的硬件结构示意图。

具体实施方式

[0032] 下面将参照附图更详细地描述本申请的实施例。虽然附图中显示了本申请的某些实施例，然而应当理解的是，本申请可以通过各种形式来实现，而且不应该被解释为限于这

里阐述的实施例，相反提供这些实施例是为了更加透彻和完整地理解本申请。应当理解的是，本申请的附图及实施例仅用于示例性作用，并非用于限制本申请的保护范围。

[0033] 本申请中使用的术语“包括”及其变形是开放性包括，即“包括但不限于”。术语“基于”是“至少部分地基于”。术语“一个实施例”表示“至少一个实施例”；术语“另一实施例”表示“至少一个另外的实施例”；术语“一些实施例”表示“至少一些实施例”。其他术语的相关定义将在下文描述中给出。

[0034] 需要注意，本申请中提及的“第一”、“第二”等概念仅用于对不同的装置、模块或单元进行区分，并非用于限定这些装置、模块或单元所执行的功能的顺序或者相互依存关系。

[0035] 需要注意，本申请中提及的“一个”、“多个”的修饰是示意性而非限制性的，本领域技术人员应当理解，除非在上下文另有明确指出，否则应该理解为“一个或多个”。

[0036] 下面从时间序列异常数据检测装置的角度本申请提供的时间序列异常数据检测方法进行说明，该时间序列异常数据检测装置可以为服务器，也可以为服务器中的服务单元，具体不做限定。

[0037] 请参阅图1，图1为本申请实施例提供的时间序列异常数据检测方法的流程示意图，包括：

[0038] 101、获取预设时长内的时间序列数据。

[0039] 本实施例中，时间序列异常数据检测装置可以获取预设时长内的时间序列数据，具体的，时间序列异常数据检测装置可以直接从数据库中拉取预设时长内的时间序列数据，该数据库中存储有多种形式以及多个时间段内的时间序列数据，该预设时长例如可以为当前时刻之前的一天，也即获取前一天的时间序列数据，当然也还可以为其他的预设时长，例如获取过去5天之内的时间序列数据，具体可以根据实际的应用场景来决定。

[0040] 102、对时间序列数据进行预处理，得到目标时间序列。

[0041] 本实施例中，时间序列异常数据监测装置在得到时间序列数据之后，可以对该时间序列数据进行预处理，最终得到目标时间序列，该预处理主要包括数据提取、缺失值处理、数据归一化，其中，数据提取是将时间序列数据的格式调整为统一的格式，调整数据时间戳，解决数据时间戳顺序不规整问题，使其按时间戳从小到大排序，并去除重复数据；数据提取后的时间序列数据包含如下属性字段：时间戳以及时间序列的值。缺失值处理，首先根据时间序列数据获取数据的采样间隔，之后在该采样间隔内获得时间序列数据的时间区间内应有的数据点个数、实际的数据点个数以及缺失的数据点个数，在已有数据点的基础上，将对应时间点缺失的数据点进行缺失值处理，使用线性插值的方式补充缺失的数据点；设经过缺失值处理后的时间序列为$X=\{X_i, i=1,2,\cdots,n\}$，其中数据归一化是对经过缺失值处理后的时间序列进行min-max归一化处理，具体的可以通过如下公式进行归一化处理：

[0042] $$Y=\frac{X-\min(X)}{\max(X)-\min(X)}。$$

[0043] 103、对目标时间序列进行周期判断，得到目标时间序列对应的目标周期。

[0044] 本实施例中，时间序列异常数据监测装置在得到目标时间序列之后，可以对该目标时间序列进行周期判断，得到目标时间序列对应的目标时间周期。具体的：时间序列异常数据检测装置可以使用自回归模型，即自相关系数/序列相关函数(Auto correlation

Function,ACF),再给定一个时间序列X,对于不同的时间延迟l,有:

[0045] $$ACF(l)=\frac{\sum_{i=1}^{n}Y_iY_{i+l}}{n},l=1,2,\cdots,\frac{n}{2};$$

[0046] 其中,n为目标时间序列中数据点的数量,Y_i为目标时间序列中的第i个序列,假设目标周期为T,自相关性在某些滞后时变得很高,比如1T,2T,3T等等。对应于ACF中第一个峰值的滞后被视为目标时间序列对应的目标周期,即目标周期T可以通过如下公式得到:

[0047] $$T=\arg\max_{l}ACF(l)$$ 。

[0048] 104、基于目标周期对目标时间序列进行分解,得到目标时间序列对应的目标输出结果。

[0049] 本实施例中,目标时间序列通常包含复杂的季节性、趋势和噪声特征,因此很难去识别异常数据点,因此时间序列异常数据检测装置可以利用时间序列分解技术,如STL分解算法以及经验模态分解算法(Empirical Mode Decomposition,EMD),在得到目标周期之后,可以基于目标周期将时序分解为季节项、趋势项和残差项。下面分别从STL分解算法以及EM分解算法对目标时间序列的分解进行说明,具体如下:

[0050] 1、STL分解。

[0051] STL(Seasonal and Trend decomposition using Loess)分解算法,是以鲁棒局部加权回归作为平滑方法的时间序列分解方法。通过STL分解算法将目标时间序列分解成季节项、趋势项和残差项,其中该季节项、趋势项和残差项为不同的时间序列,具体的可以表示$Y_n=S_n+T_n+R_n$,其中,S_n为季节项,T_n为趋势项,R_n为残差项,也就是说,目标时间序$Y_n=\{Y_i,i=1,2,\cdots,n\}$,经过STL分解后输出为三条时间序列,分别为:季节项$S_n=\{S_i,i=1,2,\cdots,n\}$,趋势项$T_n=\{T_i,i=1,2,\cdots,n\}$,残差项$R_n=\{R_i,i=1,2,\cdots,n\}$。为了便于描述,把趋势项和季节项之和记为:$A_n=S_n+T_n$。请参阅图2,图2为本申请实施例提供的STL分解效果示意图,包括目标时序序列的示意图201、趋势项的示意图202、季节项的示意图203以及残差项的示意图204。

[0052] 2、EMD分解:

[0053] 对于输入的时间序列Y_n,EMD分解算法将其分解为多条分量IMFs,取最高频的IMF1为残差项。为了便于描述,将EMD分解算法输出的最高频信号IMF1记为残差项$R_n=\{R_i,i=1,2,\cdots,n\}$,其余的IMF分量之和记为其余$A_n=\sum_{j=2}^{K}IMF_j$,其中K为EMD分解后得到的分量数量。

[0054] 由此,可以将目标时间序列经过STL分解和EMD分解,分别得到两种分解方法输出的残差项R_n和其余项A_n。

[0055] 需要说明的是,上述以STL分解和EMD分解为例进行说明,在实际应用中,当然还可以有其他的时间序列的分解方式,具体不做限定。

[0056] 105、根据目标输出结果的残差项以及其余项确定目标时间序列中的异常数据。

[0057] 本实施例中,时间序列异常数据检测装置在得到目标输出结果的残差项以及其余

项之后，可以根据该目标输出结果的残差项以及其余项确定目标时间序列中的异常数据。

[0058] 需要说明的是，时间序列异常数据检测装置主要通过统计检验的相关方法：Nsigma(其中该N的取值可以根据实际情况进行选择，例如可以为3sigma，也可以为4sigma，具体不做限定)和四分位数，目标时间序列的残差项进行异常检测，从而获得时间序列中的异常数据，同时，结合残差的统计分析结果、季节项和趋势项，确定目标时间序列的上界线、下界线和基线，如果目标时间序列中存在不在上界线和下界线之间的数据点，则认为该数据点为异常数据。

[0059] 一个实施例中，该其余项包括季节项以及趋势项，时间序列异常数据检测装置根据目标输出结果的残差项以及残差项确定目标时间序列中的异常数据包括：

[0060] 根据目标周期计算第一目标数据点的平均值，第一目标数据点为目标输出结果的残差项中任意一个数据；

[0061] 计算第一数据点与第二数据点的目标和值，第一数据点为目标输出结果的季节项与第一目标数据点对应的数据点，第二数据点为目标输出结果的趋势项中与第一目标数据点对应的数据点；

[0062] 根据第一目标数据点的平均值以及目标和值确定第一基线点、第一上界线点以及第一下界线点，其中，第一动态基线点为第一动态基线中与第一目标数据点对应的基线点，第一上界线点为第一上界线中与第一目标数据点对应的上界线点，第一下界线点为第一下界线中与第一目标数据点对应的下界线点；

[0063] 将目标时间序列中不处于第一上界线与第一下界线的数据点确定为目标时间序列的异常数据。

[0064] 本实施例中，时间序列异常数据检测装置可以设置滑动窗口长度为目标周期T，在残差项上计算平均值，该残差项上的任意一个第i个数据点的平均值M_i可以通过如下公式计算：

[0065] $$M_i=\begin{cases}\dfrac{1}{T}\sum_{i=1}^{T}R_i, i=T\\ M_{i+1}+\dfrac{(R_i-R_{i-T})}{T}, T<i\le n\end{cases};$$

[0066] 其中，R_i第一目标数据点，T为所述目标周期；

[0067] 时间序列异常数据检测装置可以根据M_i以及目标和值计算第一基线点、第一上界线点以及第一下界线点，具体的，可以通过如下公式计算锁死第一基线点：

[0068] $B_i=M_i+A_i$；

[0069] 其中，B_i为第一基线点，A_i为目标和值，$A_i=T_i+S_i$，T_i为第二数据点，S_i为第一数据点；

[0070] 通过如下公式计算第一上界线点：

[0071] $$B_i^{upper}=A_i+M_i+3\times\sigma_i;$$

[0072] 其中，B_i^{upper} 为第一上界线点，σ_i 为第一目标数据点对应的标准差，$\sigma_i = \sqrt{\frac{1}{t}\sum_{j=i-T+1}^{i}(R_j - M_j)^2}$；

[0073] 通过如下公式计算第一下界线点：

[0074] $$B_i^{lower} = A_i + M_i - 3 \times \sigma_i ;$$

[0075] 其中，B_i^{lower} 为第一下界线点。

[0076] 由此，可以得到第一动态基线、第一上界线以及第一下界线，之后将目标时间序列中不处于第一上界线与第一下界线之间的数据点确定为目标时间序列的异常数据。

[0077] 可以理解的是，在得到目标时间序列的异常数据之后，为了方便用户直观的看出时间序列中哪些数据点出现异常，可以通过图形的方式进行展示，下面结合图3以及图4对两种不同的方式得到的目标输出结果进行计算得到的第一动态基线、第一上界线以及第一下界线与目标时间序列对应的曲线进行说明：

[0078] 请参阅图3，图3为本申请实施例提供STL分解算法以及Nsigma算法相结合得到时间序列异常数据的示意图，301为第一上界线，302为第一基线，303为目标时间序列对应的曲线，304为第一下界线，在将这几条曲线结合进行显示，可以图示化的表明目标时间序列对应的异常数据点3051、3052、3053、3054、3055以及3056。

[0079] 请参阅图4，图4为本申请实施例提供的EMD分解算法以及Nsigma算法相结合得到时间序列异常数据的示意图，401为第一上界线，402为第一基线，403为目标时间序列对应的曲线，404为第一下界线，在将这几条曲线结合进行显示，可以图示化的表明目标时间序列对应的异常数据点4051以及4052。

[0080] 一个实施例中，时间序列异常数据检测装置根据目标输出结果的残差项以及其余项确定目标时间序列中的异常数据包括：

[0081] 确定第二目标数据点的第一四分位数、中位数以及第三四分位数，第二目标数据点为目标输出结果的残差项中的任意一个数据点；

[0082] 确定其余项中与第一目标数据点对应的第三数据点的值；

[0083] 根据所述中位数以及所述第三数据点确定第二基线点，所述第二基线点为第二动态基线中与所述目标数据点对应的基线点；

[0084] 根据第一四分位数、第三四分位数以及第三数据点的值计算第二上界线点以及第二下界线点，第二上界线点为第二上界线中与述目标数据点对应的上界线点，第二下界线点为第二下界线中与目标数据点对应的下界线点；

[0085] 将目标时间序列中不处于第二上界线与所述第二下界线之间的数据点确定为目标时间序列的异常数据。

[0086] 本实施例中，时间序列异常数据检测装置可以基于四分位的异常检测方法，通过目标输出结果的残差项以及其余项确定目标时间序列中的异常数据，具体的可以首先确定第二目标数据点的第一四分位数$Q_{1(i')}$、中位数$Q_{2(i')}$以及第三四分位数$Q_{3(i')}$，该第二目标数据点为目标输出结果的残差项中的任意一个数据点，也就是说，在残差项R_n上的第一四分

位数$Q_{1(i')}$、中位数$Q_{2(i')}$以及第三四分位数$Q_{3(i')}$分别为残差项R_n在区间[i-T+1]上的第一四分位数、中位数(第二四分位数)和第三四分位数，并确定其余项中与第一目标数据点对应的第三数据点$A_{i'}$的值，之后计算与第二目标数据点对应的第二基线点、第二上界线点以及第二下界线点，具体操作如下：

[0087] 通过如下公式计算第二基线点：

[0088] $B_{i'}=Q_{2(i')}+A_{i'}$；

[0089] 其中，$B_{i'}$为第二基线点，$A_{i'}$为第三数据点的值，$Q_{2(i')}$为中位数；

[0090] 通过如下公式计算第二上界线点：

[0091] $$B_{i'}^{upper}=A_{i'}+Q_{3(i')}+k\times(Q_{3(i')}-Q_{1(i')})$$；

[0092] 其中，$B_{i'}^{upper}$为第二上界线点，$Q_{3(i')}$为第三四分位数，$Q_{1(i')}$为第一四分位数，k为常数，其中，根据应用场景和数据特征不同，选定不同的常数k，本申请实施例选定k＝1.5，当然也还可以为其他的值，例如k＝2，具体不限定；

[0093] 通过如下公式计算第二下界线点：

[0094] $$B_{i'}^{lower}=A_{i'}+Q_{1(i')}-k\times(Q_{3(i')}-Q_{1(i')})$$；

[0095] 其中，$B_{i'}^{lower}$为第二下界线点。

[0096] 由此可以构建与目标时间序列对应的第二动态基线、第二上界线以及第二下界线，之后即可以通过该第二上界线、第二下界线以及第二动态基线来确定目标序列中的异常数据，也即将目标序列中不处于第二上界线与第二下界线之间的数据点确定为该目标时间序列的异常数据。

[0097] 可以理解的是，在得到目标时间序列的异常数据之后，为了方便用户直观的看出时间序列中哪些数据点出现异常，可以通过图形的方式进行展示，下面结合图5以及图6对两种不同的方式得到的目标输出结果进行计算得到的第二动态基线、第二上界线以及第二下界线与目标时间序列对应的曲线进行说明：

[0098] 请参阅图5，图5为本申请实施例提供STL分解算法以及四分位数算法相结合得到时间序列异常数据的示意图，501为第二上界线，502为第二基线，503为目标时间序列对应的曲线，504为第二下界线，在将这几条曲线结合进行显示，可以图示化的表明目标时间序列对应的异常数据点5051以及5052。

[0099] 请参阅图6，图6为本申请实施例提供的EMD分解算法以及四分位数算法相结合得到时间序列异常数据的示意图，601为第二上界线，602为第二基线，603为目标时间序列对应的曲线，604为第二下界线，在将这几条曲线结合进行显示，可以图示化的表明目标时间序列对应的异常数据点6051以及6052。

[0100] 需要说明的是，在实际应用中还可以综合两个时间序列分解方法STL和EMD以及两个异常检测方法Nsigma和四分位数，进行两两组合，可以得到四个对时间序列异常数据进行检测的方式，分别是STL+Nsigma、STL+四分位数、EMD+Nsigma以及EMD+四分位数，由此可

以得到四个结果，之后可以通过投票决定某个数据点的异常与否，例如针对某个数据点来说，如果这4种方式中有3种方式都确定该数据点是异常的，那么该数据点为时间序列中的异常数据点，也就是说只要有半数以上的方式认为该数据点为异常数据点，那么该数据点即为异常数据点，结合参阅图3至图6，可以看出，针对同一个数据点：3051、4051、5051以及6051对应的数据点，通过4种方式都认为该数据点为异常数据点，那么该数据点即为异常数据；如果说只有一种方式确定该数据点是异常的，那么该数据点就不为时间序列中的异常数据点；如果有两种方式确定该数据点是异常的，另种方式确定该数据点不是异常的，那么可以确定根据4种方式的权重来进行确定，也就是说可以对四种方式分别设置不同的权重，在根据权重来确定数据点是否为异常数据，例如STL+Nsigma以及EMD+Nsigma方式的权重高，且异常数据也是通过这两种方式检测出来的，那么就可以确定该数据点为异常数据，反之则不为异常数据。

[0101] 还需要说明的是，在实际应用中还可以不使用全部的4种方式，可以选择其中的至少1种来对时间序列异常数据的检测，例如只选择STL+Nsigma和EMD+四分位数，具体不做限定。

[0102] 综上所述，可以看出，本申请提供的实施例中，时间序列异常数据检测装置对预设时长内的时间序列数据进行预处理，并对预处理后的时间序列数据进行周期判断得到周期，并基于周期进行分解，得到时间序列的残差项和其余项，之后根据残差项和其余项确定目标时间序列中的异常数据。由此，可以不需要大量的特征工程和人工标注，也无需大量内存用于存储向量描述，减少工作量及内存成本，降低人力成本。

[0103] 可以理解的是，附图中的流程图和框图，图示了按照本申请各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，该模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0104] 本申请实施方式中的多个装置之间所交互的消息或者信息的名称仅用于说明性的目的，而并不是用于对这些消息或信息的范围进行限制。

[0105] 虽然采用特定次序描绘了各操作，但是这不应当理解为要求这些操作以所示出的特定次序或以顺序次序执行来执行。在一定环境下，多任务和并行处理可能是有利的。

[0106] 应当理解，本申请的方法实施方式中记载的各个步骤可以按照不同的顺序执行，和/或并行执行。此外，方法实施方式可以包括附加的步骤和/或省略执行示出的步骤。本申请的范围在此方面不受限制。

[0107] 另外，本申请还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申请的操作的计算机程序代码，上述程序设计语言包括但不限于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)。

[0108] 上面从的时间序列异常数据检测方法的角度对本申请进行说明，下面从时间序列异常数据检测装置的角度对本申请进行说明。

[0109] 请参阅图7，图7为本申请实施例提供的一种时间序列异常数据检测装置的虚拟结构示意图，该时间序列异常数据检测装置700包括：

[0110] 获取单元701，用于获取预设时长内的时间序列数据；

[0111] 预处理单元702，用于对所述时间序列数据进行预处理，得到目标时间序列；

[0112] 周期判断单元703，用于对所述目标时间序列进行周期判断，得到所述目标时间序列对应的目标周期；

[0113] 分解单元704，用于基于所述目标周期对所述目标时间序列进行分解，得到所述目标时间序列对应的目标输出结果，所述目标输出结果包括残差项以及其余项；

[0114] 确定单元705，用于根据所述目标输出结果的残差项以及其余项确定所述目标时间序列中的异常数据。

[0115] 一种可能的实现方式中，所述周期判断单元703具体用于：

[0116] 通过如下公式计算所述目标时间序列对应的目标周期：

[0117] $$T = \arg\max_{l} ACF(l)$$；

[0118] 其中，T为所述目标周期，ACF(l)为所述目标时间序列对应的不同的时间延迟l的自相关系数/序列相关函数，通过如下公式计算ACF(l)：

[0119] $$ACF(l) = \frac{\sum_{i=1}^{n} Y_i Y_{i+l}}{n}, l = 1, 2, \cdots, \frac{n}{2}$$；

[0120] 其中，n为所述目标时间序列中的数据点的数量，Y_i为所述目标时间序列中的第i个序列。

[0121] 一种可能的实现方式中，所述其余项包括季节项以及趋势项，所述确定单元705具体用于：

[0122] 根据所述目标周期计算第一目标数据点的平均值，所述第一目标数据点为所述目标输出结果的残差项中任意一个数据；

[0123] 计算第一数据点与第二数据点的目标和值，所述第一数据点为所述目标输出结果的季节项与所述第一目标数据点对应的数据点，所述第二数据点为所述目标输出结果的趋势项中与所述第一目标数据点对应的数据点；

[0124] 根据所述第一目标数据点的平均值以及所述目标和值确定第一基线点、第一上界线点以及第一下界线点，其中，所述第一动态基线点为第一动态基线中与所述第一目标数据点对应的基线点，所述第一上界线点为第一上界线中与所述第一目标数据点对应的上界线点，所述第一下界线点为第一下界线中与所述第一目标数据点对应的下界线点；

[0125] 将所述目标时间序列中不处于所述第一上界线与所述第一下界线之间的数据点

确定为所述目标时间序列的异常数据。

[0126] 一种可能的实现方式中，所述确定单元705根据所述目标周期计算第一目标数据点的平均值包括：

[0127] 通过如下公式计算所述第一目标数据点的平均值：

[0128] $$M_i=\begin{cases}\dfrac{1}{T}\sum_{i=1}^{T}R_i, i=T\\ M_{i+1}+\dfrac{(R_i-R_{i-T})}{T}, T<i\le n\end{cases};$$

[0129] 其中，M_i为所述第一目标数据点的平均值，R_i为所述第一目标数据点，T为所述目标周期；

[0130] 所述确定单元705根据所述第一目标数据点的平均值以及所述目标和值确定第一基线点、第一上界线点以及第一下界线点包括：

[0131] 通过如下公式计算所述第一基线点：

[0132] $B_i=M_i+A_i$；

[0133] 其中，B_i为所述第一基线点，A_i为所述目标和值，$A_i=T_i+S_i$，T_i为所述第二数据点，S_i为所述第一数据点；

[0134] 通过如下公式计算所述第一上界线点：

[0135] $$B_i^{upper}=A_i+M_i+3\times\sigma_i;$$

[0136] 其中，B_i^{upper}为所述第一上界线点，σ_i为所述第一目标数据点对应的标准差，

$$\sigma_i=\sqrt{\frac{1}{t}\sum_{j=i-T+1}^{i}(R_j-M_j)^2};$$

[0137] 通过如下公式计算所述第一下界线点：

[0138] $$B_i^{lower}=A_i+M_i-3\times\sigma_i;$$

[0139] 其中，B_i^{lower}为所述第一下界线点。

[0140] 一种可能的实现方式中，所述确定单元705还具体用于：

[0141] 确定第二目标数据点的第一四分位数、中位数以及第三四分位数，所述第二目标数据点为所述目标输出结果的残差项中的任意一个数据点；

[0142] 确定所述其余项中与所述第一目标数据点对应的第三数据点的值；

[0143] 根据所述中位数以及所述第三数据点确定第二基线点，所述第二基线点为第二动态基线中与所述目标数据点对应的基线点；

[0144] 根据所述第一四分位数、所述第三四分位数以及所述第三数据点计算第二上界线点以及第二下界线点，所述第二上界线点为第二上界线中与所述目标数据点对应的上界线点，所述第二下界线点为第二下界线中与所述目标数据点对应的下界线点；

[0145] 将所述目标时间序列中不处于所述第二上界线与所述第二下界线之间的数据点确定为所述目标时间序列的异常数据。

[0146] 一种可能的实现方式中，所述确定单元705根据所述中位数以及所述第三数据点确定第二基线点包括：

[0147] 通过如下公式计算所述第二基线点：

[0148] $B_{i'}=Q_{2(i')}+A_{i'}$；

[0149] 其中，$B_{i'}$为所述第二基线点，$A_{i'}$为所述第三数据点，$Q_{2(i')}$为所述中位数；

[0150] 所述确定单元705根据所述第一四分位数、所述第三四分位数以及所述第三数据点计算第二上界线点以及第二下界线点包括：

[0151] 通过如下公式计算所述第二上界线点：

[0152] $$B_{i'}^{upper}=A_{i'}+Q_{3(i')}+k\times(Q_{3(i')}-Q_{1(i')})\ ;$$

[0153] 其中，$B_{i'}^{upper}$为所述第二上界线点，$Q_{3(i')}$为所述第三四分位数，$Q_{1(i')}$为所述第一四分位数，k为常数；

[0154] 通过如下公式计算所述第二下界线点：

[0155] $$B_{i'}^{lower}=A_{i'}+Q_{1(i')}-k\times(Q_{3(i')}-Q_{1(i')})\ ;$$

[0156] 其中，$B_{i'}^{lower}$为所述第二下界线点。

[0157] 综上所述，可以看出，本申请提供的实施例中，时间序列异常数据检测装置对预设时长内的时间序列数据进行预处理，并对预处理后的时间序列数据进行周期判断得到周期，并基于周期进行分解，得到时间序列的残差项和其余项，之后根据残差项和其余项确定目标时间序列中的异常数据。由此，可以不需要大量的特征工程和人工标注，也无需大量内存用于存储向量描述，减少工作量及内存成本，降低人力成本。

[0158] 需要说明的是，描述于本申请实施例中所涉及到的单元可以通过软件的方式实现，也可以通过硬件的方式来实现。其中，单元的名称在某种情况下并不构成对该单元本身的限定，例如，获取单元还可以被描述为“获取目标用户的证件信息的单元”。

[0159] 本文中以上描述的功能可以至少部分地由一个或多个硬件逻辑部件来执行。例如，非限制性地，可以使用的示范类型的硬件逻辑部件包括：现场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、片上系统(SOC)、复杂可编程逻辑设备(CPLD)等等。

[0160] 请参阅图8，图8为本申请实施例提供的一种机器可读介质的实施例示意图。

[0161] 如图8所示，本实施例提供了一种机器可读介质800，其上存储有计算机程序811，该计算机程序811被处理器执行时实现上述图1所述时间序列异常数据检测方法的步骤。

[0162] 需要说明的是，本申请的上下文中，机器可读介质可以是有形的介质，其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备，或者上述

内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或快闪存储器)、光纤、便捷式紧凑盘只读存储器(CD-ROM)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

[0163] 需要说明的是,本申请上述的机器可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本申请中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本申请中,计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读信号介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于:电线、光缆、RF(射频)等等,或者上述的任意合适的组合。

[0164] 上述计算机可读介质可以是上述电子设备中所包含的;也可以是单独存在,而未装配入该电子设备中。

[0165] 请参阅图9,图9是本申请实施例提供的一种服务器的硬件结构示意图,该服务器900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中央处理器(central processing units,CPU)922(例如,一个或一个以上处理器)和存储器932,一个或一个以上存储应用程序940或数据944的存储介质930(例如一个或一个以上海量存储设备)。其中,存储器932和存储介质930可以是短暂存储或持久存储。存储在存储介质930的程序可以包括一个或一个以上模块(图示没标出),每个模块可以包括对服务器中的一系列指令操作。更进一步地,中央处理器922可以设置为与存储介质930通信,在服务器900上执行存储介质930中的一系列指令操作。

[0166] 服务器900还可以包括一个或一个以上电源926,一个或一个以上有线或无线网络接口950,一个或一个以上输入输出接口958,和/或,一个或一个以上操作系统941,例如Windows ServerTM,Mac OS XTM,UnixTM,LinuxTM,FreeBSDTM等等。

[0167] 上述实施例中由时间序列异常数据检测装置所执行的步骤可以基于该图9所示的服务器结构。

[0168] 还需要说明的,根据本申请的实施例,上述图1中的流程示意图描述的所述时间序列异常数据检测方法的过程可以被实现为计算机软件程序。例如,本申请的实施例包括一种计算机程序产品,其包括承载在非暂态计算机可读介质上的计算机程序,该计算机程序包含用于执行上述图2的流程示意图中所示的方法的程序代码。

[0169] 尽管已经采用特定于结构特征和/或方法逻辑动作的语言描述了本主题,但是应

当理解所附权利要求书中所限定的主题未必局限于上面描述的特定特征或动作。相反，上面所描述的特定特征和动作仅仅是实现权利要求书的示例形式。

[0170] 虽然在上面论述中包含了若干具体实现细节，但是这些不应当被解释为对本申请的范围的限制。在单独的实施例的上下文中描述的某些特征还可以组合地实现在单个实施例中。相反地，在单个实施例的上下文中描述的各种特征也可以单独地或以任何合适的子组合的方式实现在多个实施例中。

[0171] 以上描述仅为本申请的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本申请中所涉及的公开范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离上述公开构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本申请中公开的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。

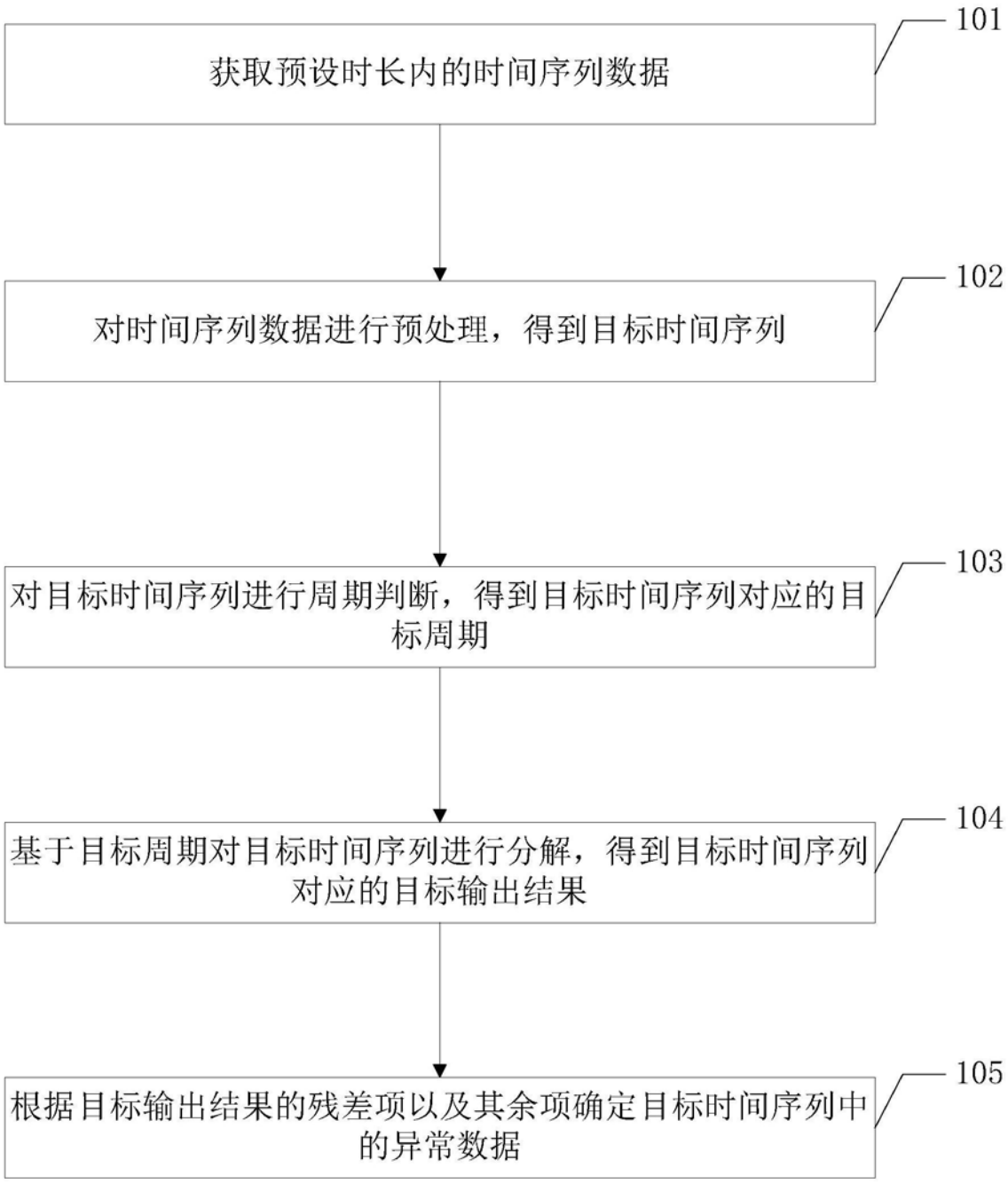

101
获取预设时长内的时间序列数据
102
对时间序列数据进行预处理，得到目标时间序列
103
对目标时间序列进行周期判断，得到目标时间序列对应的目标周期
104
基于目标周期对目标时间序列进行分解，得到目标时间序列对应的目标输出结果
105
根据目标输出结果的残差项以及其余项确定目标时间序列中的异常数据

图1

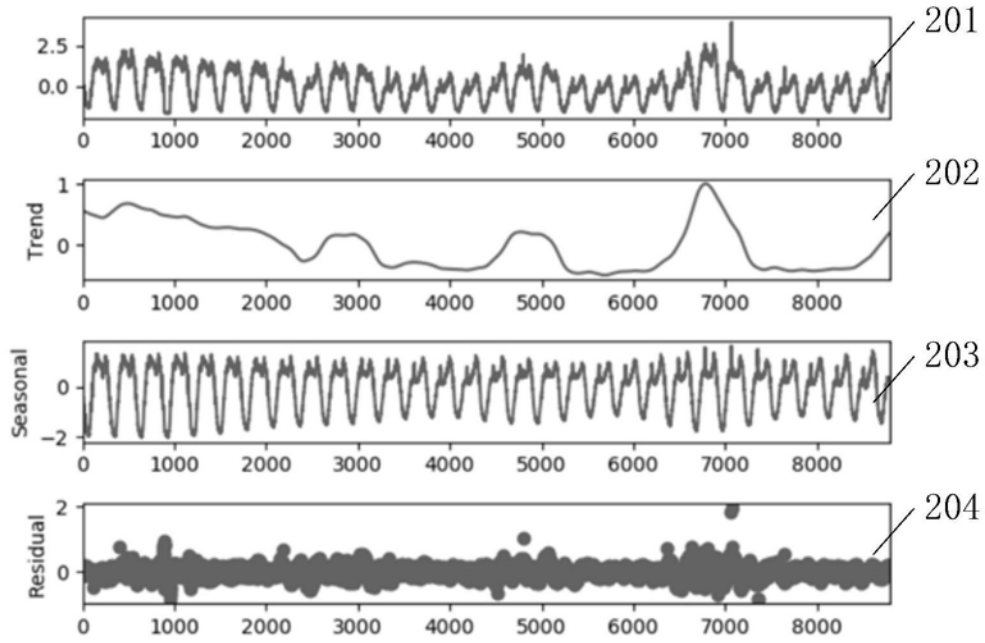

201
202
203
204
Trend
Seasonal
Residual

图2

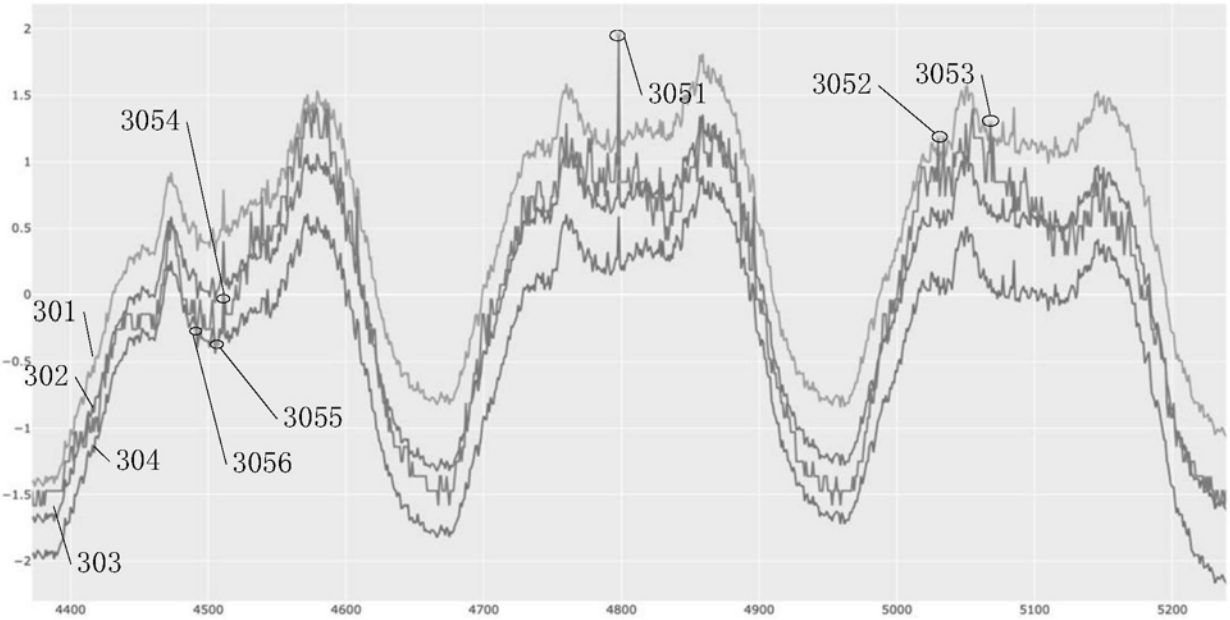

3051
3052
3053
3054
3055
3056
301
302
303
304

图3

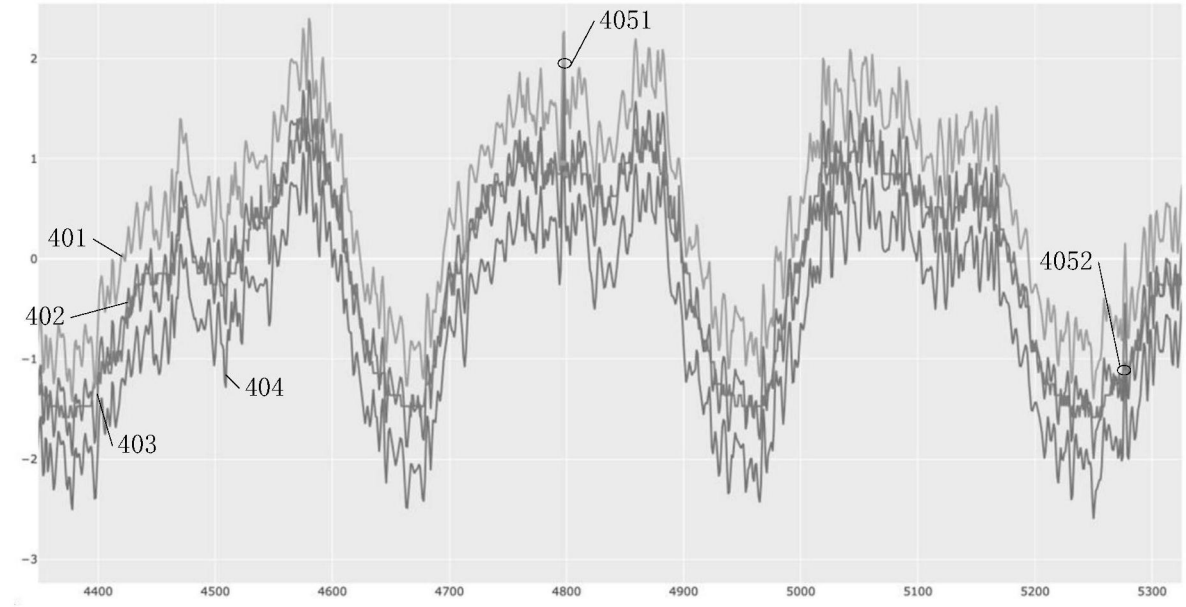

4051
401
402
404
403
4052
2
1
0
-1
-2
-3
4400
4500
4600
4700
4800
4900
5000
5100
5200
5300

图4

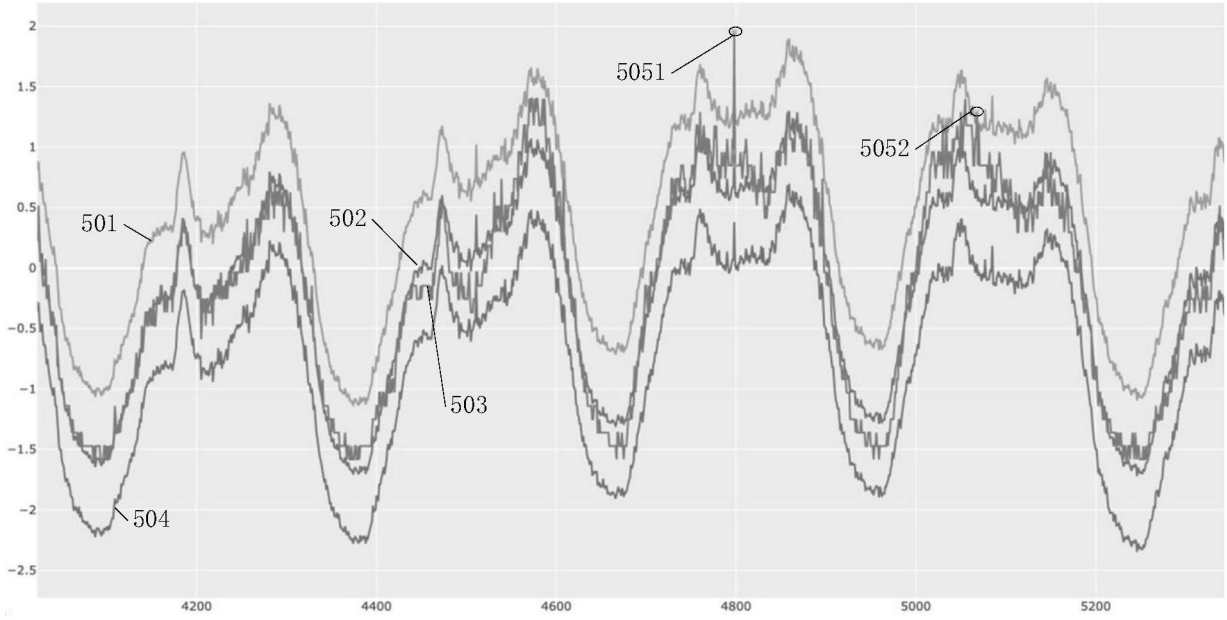

5051
5052
501
502
503
504
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
4200
4400
4600
4800
5000
5200

图5

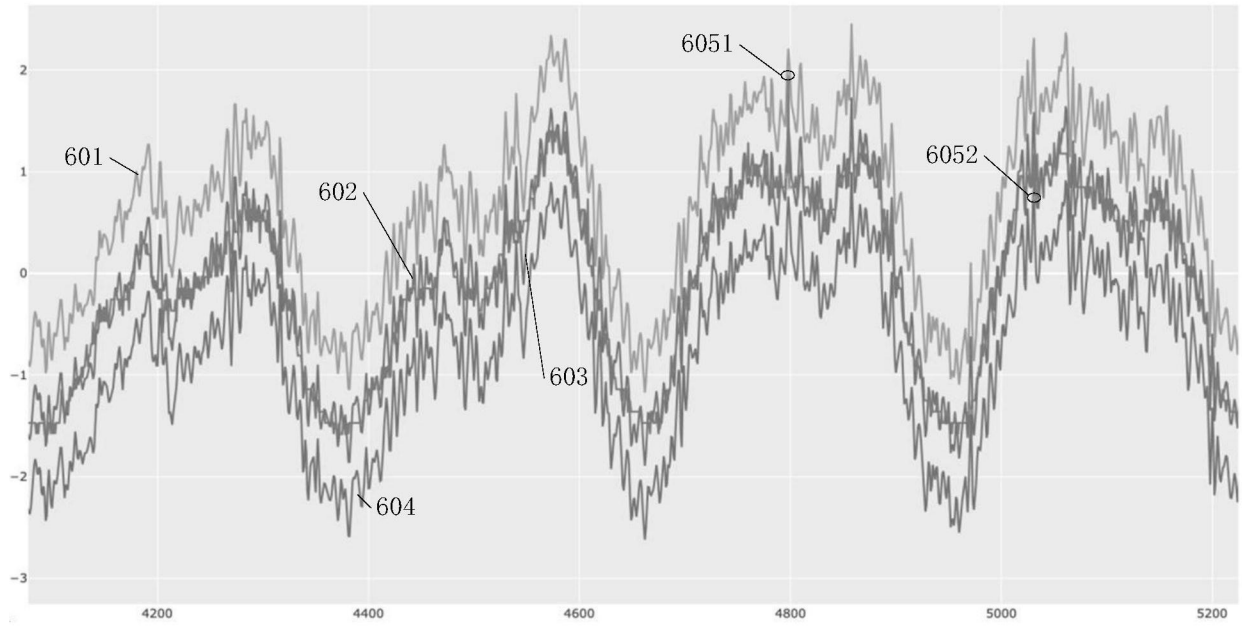

6051
601
602
6052
603
604
2
1
0
−1
−2
−3
4200
4400
4600
4800
5000
5200

图6

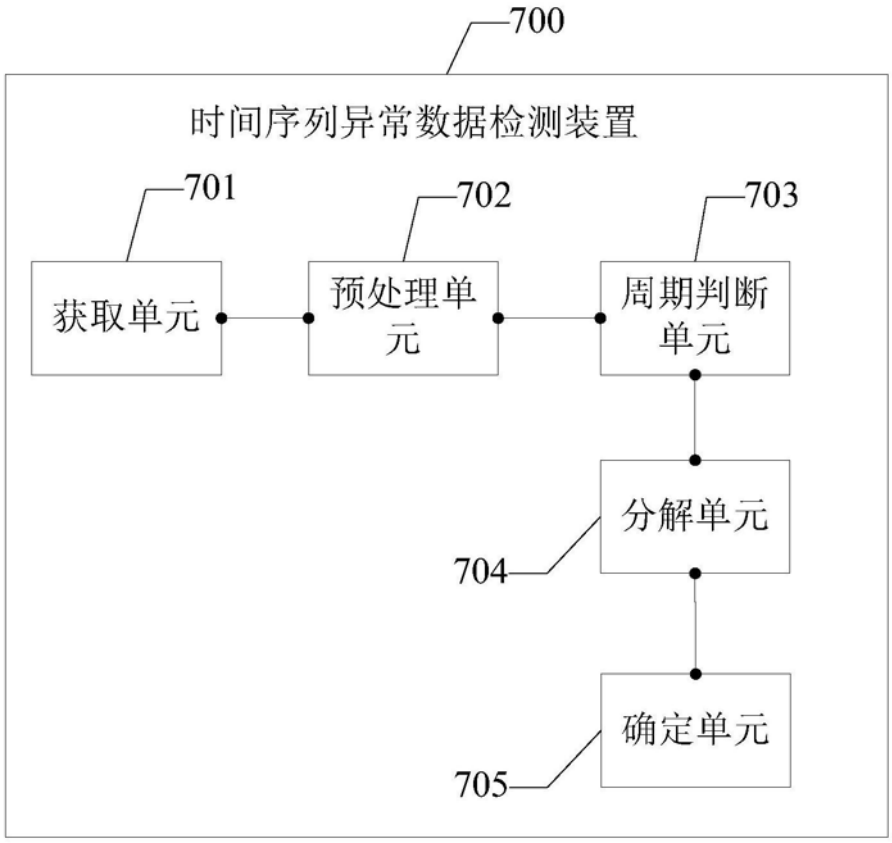

700
时间序列异常数据检测装置
701
702
703
获取单元
预处理单元
周期判断单元
分解单元
704
确定单元
705

图7

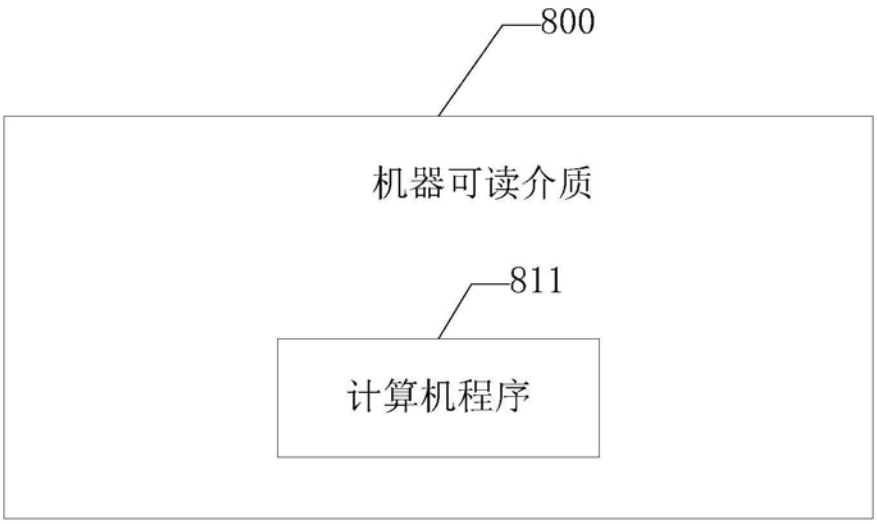

800
机器可读介质
811
计算机程序

图8

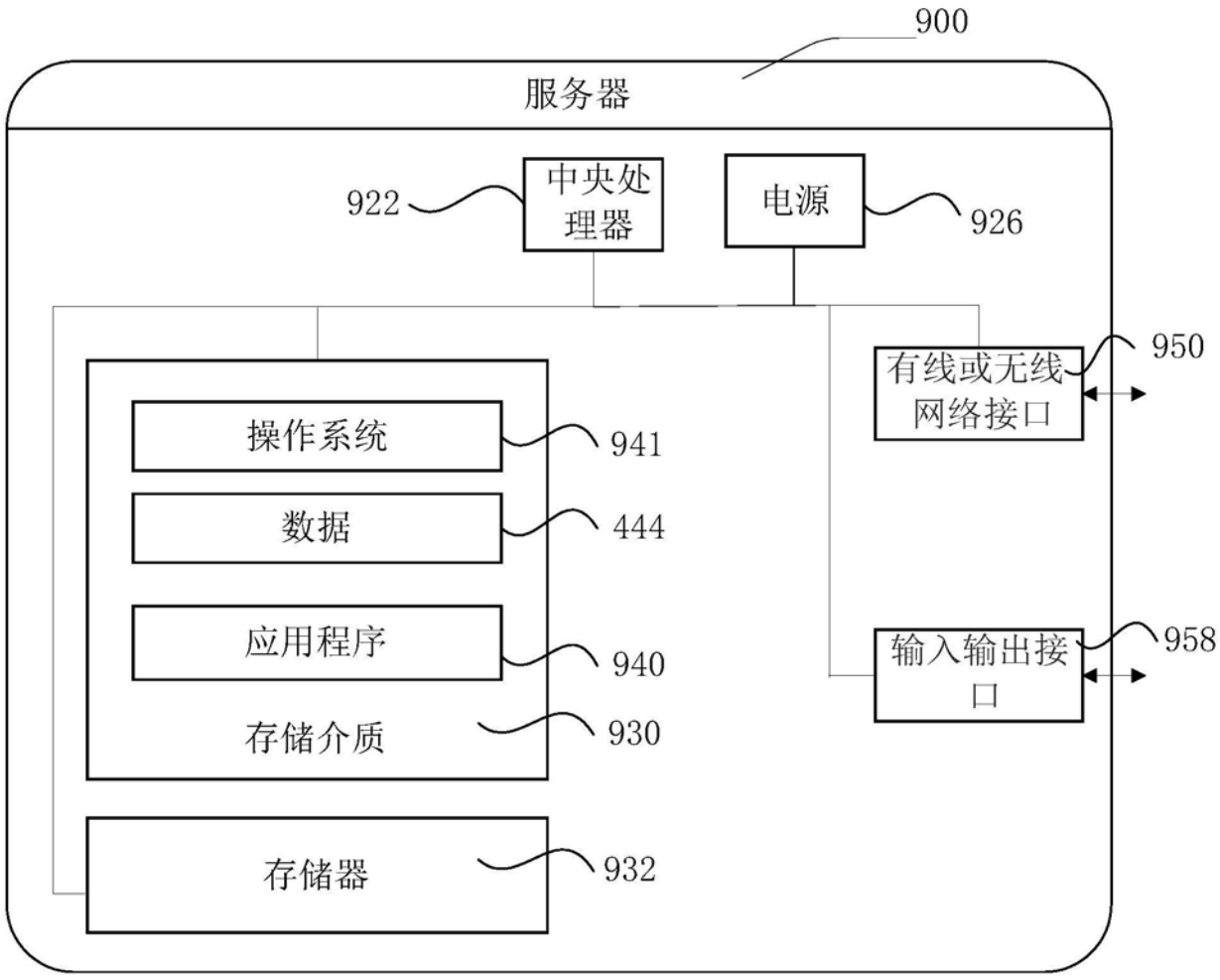

900
服务器
922
中央处理器
电源
926
950
有线或无线网络接口
操作系统
941
数据
444
应用程序
940
存储介质
930
输入输出接口
958
存储器
932

图9