(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115118511 A
(43)申请公布日 2022.09.27

(21)申请号 202210759958.3

(22)申请日 2022.06.29

(71)申请人 北京奇艺世纪科技有限公司
地址 100080 北京市海淀区北一街2号鸿城拓展大厦10、11层

(72)发明人 潘天宇 季素润

(74)专利代理机构 北京华夏泰和知识产权代理有限公司 11662
专利代理师 杜欣

(51)Int.Cl.
H04L 9/40(2022.01)
H04L 41/16(2022.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

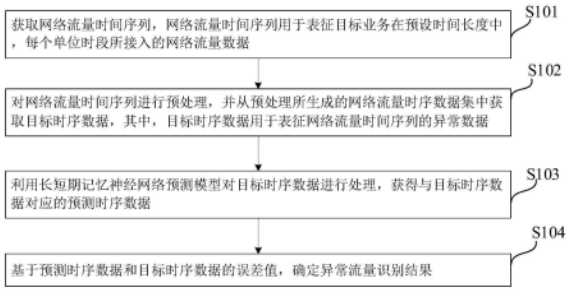
权利要求书3页 说明书12页 附图3页

(54)发明名称

异常流量识别方法、装置、电子设备及存储介质

(57)摘要

本申请涉及一种异常流量识别方法、装置、电子设备及存储介质，该方法包括：获取网络流量时间序列，网络流量时间序列用于表征目标业务在预设时间长度中，每个单位时段所接入的网络流量数据；对网络流量时间序列进行预处理，从预处理所生成的网络流量时序数据集中获取目标时序数据，目标时序数据用于表征网络流量时间序列的异常数据；利用长短期记忆神经网络预测模型对目标时序数据进行处理，获得与目标时序数据对应的预测时序数据；基于预测时序数据和目标时序数据的误差值，确定异常流量识别结果。通过本申请，解决异常流量识别困难、误报率高的问题。



CN 115118511 A

1.一种异常流量识别方法,其特征在于,包括:

获取网络流量时间序列,其中,所述网络流量时间序列用于表征目标业务在预设时间长度中,每个单位时段所接入的网络流量数据;

对所述网络流量时间序列进行预处理,并从预处理所生成的网络流量时序数据集中获取目标时序数据,其中,所述目标时序数据用于表征所述网络流量时间序列的异常数据;

利用长短期记忆神经网络预测模型对所述目标时序数据进行处理,获得与所述目标时序数据对应的预测时序数据;

基于所述预测时序数据和所述目标时序数据的误差值,确定异常流量识别结果。

2.根据权利要求1所述的方法,其特征在于,获取网络流量时间序列,包括:

从预设文件系统中获取对应的网络流量数据;

将所述网络流量数据转化预设的键值对key-value格式,得到所述网络流量时间序列,其中,所述key-value中key用于表征一个所述目标业务,所述key-value中value用于表征目标业务在预设时间长度中,每个单位时段接入的网络流量数据。

3.根据权利要求2所述的方法,其特征在于,对所述网络流量时间序列进行预处理,包括:

检测所述网络流量时间序列的缺失值;

基于所述缺失值,对所述网络流量时间序列进行过滤处理,得到候选网络流量时间序列;

根据所述候选网络流量时间序列,生成多种类别的第一流量时序数据,其中,每种所述第一流量时序数据对应一种设定的变化周期;

过滤多种所述第一流量时序数据中变化周期小于流量变化周期阈值的第一流量时序数据,并基于过滤得到的第二流量时序数据,生成所述网络流量时序数据集。

4.根据权利要求3所述的方法,其特征在于,根据所述候选网络流量时间序列,生成多种类别的第一流量时序数据,包括:利用时间序列分解算法STL对所述候选网络流量时间序列进行分解,生成多种所述第一流量时序数据。

5.根据权利要求4所述的方法,其特征在于,基于过滤得到的第二流量时序数据,生成所述网络流量时序数据集,包括:

利用移动窗递推算法,对所有所述第二流量时序数据进行加窗,得到每条所述第二流量时序数据所对应的子时序数据集,其中,所述网络流量时序数据集包括每条所述第二流量时序数据所对应的子时序数据集。

6.根据权利要求5所述的方法,其特征在于,所述第二流量时序数据至少包括残差项时序数据,从预处理所生成的网络流量时序数据集中获取目标时序数据,包括:

从所述网络流量时序数据集所对应的多种类别的所述第二流量时序数据中,选取所述残差项时序数据;

将所述残差项时序数据作为所述目标时序数据,其中,所述目标时序数据包括每条所述残差项时序数据对应的子时序数据集。

7.根据权利要求6所述的方法,其特征在于,基于所述预测时序数据和所述目标时序数据的误差值,确定异常流量识别结果,包括:

确定所述残差项时序数据所对应多个所述误差值,其中,所述误差值为所述残差项时

序数据与所对应的所述预测时序数据的差值的绝对值；

判断多个所述误差值之和是否大于预设误差值阈值；

在判断到多个所述误差值之和大于所述预设误差值阈值的情况下，确定所述网络流量时间序列所对应的网络流量数据为异常网络流量数据。

8.根据权利要求3所述的方法，其特征在于，基于所述缺失值，对所述网络流量时间序列进行过滤处理，得到候选网络流量时间序列，包括：

删除所述缺失值大于缺失值阈值的网络流量时间序列，得到所述候选网络流量时间序列，和/或，

利用滑动平均算法对所述缺失值进行插值填充，得到所述候选网络流量时间序列。

9.根据权利要求1所述的方法，其特征在于，所述长短期记忆神经网络预测模型包括基于注意力的长短期记忆网络Attention-based LSTM，所述长短期记忆神经网络预测模型的训练过程包括：

获取历史网络流量数据，并将所述网络流量数据按键值对key-value格式转换为对应的历史网络流量时间序列；

对所述历史网络流量时间序列进行对应的预处理，得到历史网络流量时间序列所对应的残差项时间序列，其中，所述预处理包括缺失值过滤、插值、STL分解；

对所述残差项时间序列进行增窗，得到对应的初始全量数据集，并按预设分配比将所述初始全量数据集拆解为训练集和测试集；

通过所述训练集和测试集对初始Attention-based LSTM进行训练，得到所述初始全量数据集的权重和对应的待选预测模型；

基于所述初始全量数据集的权重，对所述初始全量数据集进行加权，得到备选全量数据集；

在所述备选全量数据集中，按权重由大到小顺序选取预设数据量的数据，得到目标全量数据，并通过所述目标全量数据对所述待选预测模型进行重训练，得到所述长短期记忆神经网络预测模型。

10.一种异常流量识别装置，其特征在于，包括：

获取模块，用于获取网络流量时间序列，其中，所述网络流量时间序列用于表征目标业务在预设时间长度中，每个单位时段所接入的网络流量数据；

预处理模块，用于对所述网络流量时间序列进行预处理，并从预处理所生成的网络流量时序数据集中获取目标时序数据，其中，所述目标时序数据用于表征所述网络流量时间序列的异常数据；

处理模块，用于利用长短期记忆神经网络预测模型对所述目标时序数据进行处理，获得与所述目标时序数据对应的预测时序数据；

确定模块，用于基于所述预测时序数据和所述目标时序数据的误差值，确定异常流量识别结果。

11.一种电子设备，其特征在于，包括处理器、通信接口、存储器和通信总线，其中，处理器，通信接口，存储器通过通信总线完成相互间的通信；

存储器，用于存放计算机程序；

处理器，用于执行存储器上所存放的程序时，实现权利要求1-9任一项所述的异常流量

识别方法的步骤。

12.一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,所述计算机程序被处理器执行时实现如权利要求1-9任一项所述的异常流量识别方法的步骤。

异常流量识别方法、装置、电子设备及存储介质

技术领域

[0001] 本申请涉及智能分析技术领域，尤其涉及一种异常流量识别方法、装置、电子设备及存储介质。

背景技术

[0002] 随着计算机网络的日益发展，网络规模扩大，创新性的应用模式和需求的不断涌现，使的相关技术中对异常流量的监控越来越困难。

[0003] 相关技术中，对接入相关系统的业务的流量的监控采用基于简单的规则作流量监控，但因业务流量具有：呈现振荡状、流量变化周期性不尽相同及业务数量庞大，存在无法捕捉部分异常激增的流量的情况，同时，相关技术中的异常流量的监测，存在较高的误报率，易将正常的流量波动识别为异常变化的流量。

[0004] 针对相关技术中异常流量识别困难、误报率高的问题，尚未存在有效的解决方案。

发明内容

[0005] 本申请提供了一种异常流量识别方法、装置、电子设备及存储介质，以至少解决相关技术中异常流量识别困难、误报率高的问题。

[0006] 第一方面，本申请提供了一种异常流量识别方法，包括：获取网络流量时间序列，其中，所述网络流量时间序列用于表征目标业务在预设时间长度中，每个单位时段所接入的网络流量数据；对所述网络流量时间序列进行预处理，并从预处理所生成的网络流量时序数据集中获取目标时序数据，其中，所述目标时序数据用于表征所述网络流量时间序列的异常数据；利用长短期记忆神经网络预测模型对所述目标时序数据进行处理，获得与所述目标时序数据对应的预测时序数据；基于所述预测时序数据和所述目标时序数据的误差值，确定异常流量识别结果。

[0007] 第二方面，本申请提供了一种异常流量识别装置，包括：

[0008] 获取模块，用于获取网络流量时间序列，其中，所述网络流量时间序列用于表征目标业务在预设时间长度中，每个单位时段所接入的网络流量数据；

[0009] 预处理模块，用于对所述网络流量时间序列进行预处理，并从预处理所生成的网络流量时序数据集中获取目标时序数据，其中，所述目标时序数据用于表征所述网络流量时间序列的异常数据；

[0010] 处理模块，用于利用长短期记忆神经网络预测模型对所述目标时序数据进行处理，获得与所述目标时序数据对应的预测时序数据；

[0011] 确定模块，用于基于所述预测时序数据和所述目标时序数据的误差值，确定异常流量识别结果。

[0012] 第三方面，提供了一种电子设备，包括处理器、通信接口、存储器和通信总线，其中，处理器，通信接口，存储器通过通信总线完成相互间的通信；

[0013] 存储器，用于存放计算机程序；

[0014] 处理器，用于执行存储器上所存放的程序时，实现第一方面任一项实施例所述的异常流量识别方法的步骤。

[0015] 第四方面，提供了一种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第一方面任一项实施例所述的异常流量识别方法的步骤。

[0016] 与相关技术相比，本实施例中提供了异常流量识别方法、装置、服务器及存储介质，通过获取网络流量时间序列，网络流量时间序列用于表征目标业务在预设时间长度中，每个单位时段所接入的网络流量数据；对网络流量时间序列进行预处理，并从预处理所生成的网络流量时序数据集中获取目标时序数据，目标时序数据用于表征网络流量时间序列的异常数据；利用长短期记忆神经网络预测模型对目标时序数据进行处理，获得与目标时序数据对应的预测时序数据；基于预测时序数据和目标时序数据的误差值，确定异常流量识别结果，解决相关技术中异常流量识别困难、误报率高的问题，实现了降低业务流量异常监测误报率、减少因异常流量误报所造成的客诉及降低人力成本的有益效果。

[0017] 本申请的一个或多个实施例的细节在以下附图和描述中提出，以使本申请的其他特征、目的和优点更加简明易懂。

附图说明

[0018] 此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本发明的实施例，并与说明书一起用于解释本发明的原理。

[0019] 为了更清楚地说明本发明实施例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简单地介绍，显而易见地，对于本领域普通技术人员而言，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0020] 图1为根据本申请实施例的一种异常流量识别方法的流程示意图；

[0021] 图2是根据本申请优选实施例的训练Attention-based LSTM的流程示意图；

[0022] 图3是根据本申请实施例的一种异常流量识别装置的结构框图；

[0023] 图4是本申请实施例的电子设备的结构示意图。

具体实施方式

[0024] 为使本申请实施例的目的、技术方案和优点更加清楚，下面将结合本申请实施例中的附图，对本申请实施例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申请的一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0025] 在对本申请实施例进行描述之前，对本申请实施例的所涉及的相关技术进行说明如下：

[0026] 时间序列分解算法(Seasonal-Trend decomposition procedure based on Loess，简称STL)，是一种基于局部线性回归的季节性趋势分解算法，可以把时间序列分解为趋势项、季节项和剩余项；STL包含一系列局部加权回归平滑器，计算速度比较快，可以应对非常大的时间序列数据；STL把数据原始序列分解成长期时间趋势Trend、季节性时间趋势Seasonal，周期性时间趋势component以及剩余的残差项Remainder，长期时间趋势是流量在长期来看上涨/下降的趋势，季节性指流量在某个时间段上的突变，周期性指流量以周

期规律突变,残差项往往代表异常。

[0027] 长短期记忆网络(LongShort-TermMemory,LSTM)作为深度学习循环神经网络(Recurrent Neural Network,RNN)的一种变种,广泛应用于语音识别、自然语言处理、图像压缩、时间序列处理等序列模型处理任务中。LSTM通过引入门控机制和用于存储长短期历史信息的状态值,有效解决RNN训练过程中出现的梯度爆炸和梯度消失问题;相关技术中,打破了传统编码器-解码器结构在编解码时都依赖于内部一个固定长度向量的限制,提成了采用Attention机制,同时,将Attention机制引入LSTM,通过保留LSTM编码器对输入序列的中间输出结果,然后训练一个模型来对这些输入进行选择性的学习并且在模型输出时将输出序列与之进行关联,进而提出了结合注意力机制的长短期记忆网络(Attention-Based Long Short-Term Memory,Attention-based LSTM),在本申请中,采用Attention-based LSTM对流量时间序列的进行预测,籍以实现异常流量的识别。

[0028] 滑动/移动平均(Moving Average,简称MA),又称移动平均线,是技术分析中一种分析时间序列的常用工具,常见的移动平均包括简单移动平均(Simple Moving Average,SMA)、权重移动平均(Weight Moving Average,WMA)、指数移动平均(Exponential Moving Average,EMA)等。

[0029] 移动窗口函数(Moving Window Functions),使用滑窗(sliding windown)或呈指数降低的权重(exponentially decaying weights),来对时间序列进行统计值计算和其他一些函数计算,例如:消除噪声;移动窗口函数包括固定长度的移动窗口(moving window withfixed length)、指数加权移动平均数(exponentially weighted moving average);在本申请中,使用固定长度的移动窗口(moving window with fixed length)对相应的时间序列进行处理。

[0030] 衡量偏差标准包括均方根误差(Root Mean Square Error,简称RMSE)、均方误差(Mean Square Error,简称MSE)、平均绝对误差(Mean Absolute Error,简称MAE)以及标准差(Standard Deviation,简称SD),其中,RMSE是衡量观测值与真实值之间的偏差,常用于机器学习模型预测结果衡量的标准,MSE是真实值与预测值的差值的平方然后求和平均,常用于线性回归的损失函数,MAE是绝对误差的平均值,可以更好地反映预测值误差的实际情况,SD是方差的算术平均根,用于衡量一组数值的离散程度;本申请中采用MAE判断网络流量是否为异常流量。

[0031] 以下对本申请需要解决的相关技术中异常流量识别困难、误报率高参数的原因进行说明如下:在异常流量识别中,存在以下三种情况造成异常流量识别困难,第一种,相当一部分业务的流量呈现振荡状,且一天内的流量变化数量级较大,例如:登陆鉴权系列业务的流量在晚上18时会有激增,这对应了用户在吃晚饭的时候更倾向于登陆网站观影,而其余时间这项业务的流量较缓;第二种,每个业务的流量变化周期性不尽相同,有存在以天为周期的周期性,也有以周为周期的周期性,也存在难以用直觉捕捉的周期性,因此很难用经验或者人力去遍历这些可能性;第三种,业务数量庞大,即数据体量大,需要一个高效的方法对数据进行学习。

[0032] 对于上述三种情形,本申请通过采用STL对每项业务的流量时间序列进行过滤,得到目标时序数据(残差项时间序列),再使用Attention-based LSTM对训练样本进行加权,并基于加权的数据重新训练Attention-based LSTM,最后利用训练完备的Attention-

based LSTM以目标时序数据为输入对流量值进行预测与预警。

[0033] 本申请中描述的各种技术可用于异常流量识别的预测与检测等。

[0034] 下面将结合本申请实施例中的附图,对本申请实施例中的技术方案进行描述。

[0035] 图1为根据本申请实施例的一种异常流量识别方法的流程示意图。图1所示,本申请实施例提供了一种异常流量识别方法,该方法包括如下步骤:

[0036] 步骤S101,获取网络流量时间序列,其中,网络流量时间序列用于表征目标业务在预设时间长度中,每个单位时段所接入的网络流量数据。

[0037] 在本实施例中,获取的网络流量时间序列所对应的网络流量数据是指获取一段时间内的流量数据,例如,1天内、1周内或者1月内的目标业务所接入的流量数据。

[0038] 在本实施例,获取的网络流量时间序列对应是网络流量数据是从分布式文件系统(Hadoop Distributed File System,简称HDFS)所读取的,并在读取对应的网络流量数据后,将对应的网络流量数据进行整理,以将网络流量数据转换为预设格式(例如:python 字典),具体的,可以采用键值对key-value格式,并且,每个key代表了一个具体业务,key对应的value代表了这个业务在预设时间长度(例如:两个月)中,单位时段(例如:每个小时)接入的网络流量数据(也就是对应的流量值),然后按list形式存储网络流量时间序列。

[0039] 步骤S102,对网络流量时间序列进行预处理,并从预处理所生成的网络流量时序数据集中获取目标时序数据,其中,目标时序数据用于表征网络流量时间序列的异常数据。

[0040] 在本实施例中,在获取到网络流量时间序列之后,会对网络流量时间序列进行处理,籍以降低输入到对应的神经网络预测模型的数据对异常流量的预测所造成的影响;在本实施例中,对网络流量时间序列进行的预处理包括过滤掉丢失值过多的网络流量时间序列、对数据写入过程中造成数据丢失的网络流量时间序列进行插值处理、对网络流量时间序列进行基于局部线性回归的季节性趋势(STL)分解,并过滤掉流量变化周期小于设定变化周期的流量时间序列、对STL分解的时序数据(例如:代表异常的残差项时序数据)采用固定长度的移动窗口进行增窗;当然,获取用于训练对应的预测模型的训练数据时,也采用对预设的网络流量时间序列(例如:采集的历史的网络流量时间序列)进行对应的预处理,从而获得的对应的样本数据。

[0041] 在本实施例中,在对网络流量时间序列进行预处理,得到对应的网络流量时序数据集后,还会从网络流量时序数据集中选取对应的目标时序数据,该目标时序数据是用于预测网络流量时间序列是否为异常流量,具体地,目标时序数据为STL分解后的残差项时序数据。

[0042] 步骤S103,利用长短期记忆神经网络预测模型对目标时序数据进行处理,获得与目标时序数据对应的预测时序数据,其中,长短期记忆神经网络预测模型被训练为用于根据输入的时序数据得到与该时序数据对应的预测数据。

[0043] 在本实施例中,长短期记忆神经网络预测模型是预测前训练完备的预测模型;在其中一些可选实施方式中,用于训练该预测模型的样本数据,在训练过程中所输出的误差值可以直接用于对应采集的网络流量数据是否为异常流量的计算。

[0044] 在本实施例中,预测时序数据,是基于目标时序数据在完成预测后输出的预测值;在本实施例中,采用目标时序数据中多项时序数据预测在该多项时序数据之后的某项时序数据,例如:在目标时序数据为残差项时序数据,且在预处理过程中增窗的窗口设置设M时,

则采用第T项、第T-1、...、第T-M项的残差项时序数据预测第T+1项残差项时序数据。

[0045] 步骤S104,基于预测时序数据和目标时序数据的误差值,确定异常流量识别结果。

[0046] 在本实施例中,在完成预测某项时序数据的预测值后,会根据预测的某项时序数据的预测值与该项时序数据的实际值(预处理后输入到预测模型的目标时序数据对应的值)的误差值,也就是预测值与实际值的差值的绝对值的平均值MAE,判断对应的网络流量数据是否为异常流量。

[0047] 在其中一些可选实施方式中,对应的预警机制为:(实际值-预测值)＞15*MAE或(实际值-预测值)＜-30*MAE。

[0048] 需要说明的是,在本实施例中,异常流量识别包括流量本身的变化和流量数量的变化,流量本身的变化是指流量相关的参数、数据发生变化,例如:业务请求来源的渠道,运营商A比例多少、运营商B比例多少,在某个时间段远远超出了平均值,则属于异常流量,又例如:流量来源的城市,某个时间段大量流量来源于同一个城市,则也属于异常流量;而流量数量上的变化,则是流量数据一段时间内的流量变化,例如,某IP对应的客户端通常情况下在早8点至9点之间网络流量数据量为5GB,如果某天该IP对应的客户端在早8点至9点之间网络流量数据量为30GB,远大于通常的5GB,则认为30GB的网络流量数据量可能存在异常。

[0049] 通过步骤S101至步骤S104,采用获取网络流量时间序列,网络流量时间序列用于表征目标业务在预设时间长度中,每个单位时段所接入的网络流量数据;对网络流量时间序列进行预处理,并从预处理所生成的网络流量时序数据集中获取目标时序数据,目标时序数据用于表征网络流量时间序列的异常数据;利用长短期记忆神经网络预测模型对目标时序数据进行处理,获得与目标时序数据对应的预测时序数据;基于预测时序数据和目标时序数据的误差值,确定异常流量识别结果,解决相关技术中异常流量识别困难、误报率高的问题,实现了降低业务流量异常监测误报率、减少因异常流量误报所造成的客诉及降低人力成本的有益效果。

[0050] 在其中一些实施例中,步骤S101中的获取网络流量时间序列,可以通过如下步骤实现:

[0051] 步骤21、从预设文件系统中获取对应的网络流量数据。

[0052] 在本实施例,获取的网络流量时间序列对应是网络流量数据是从HDFS所读取的,并在读取对应的网络流量数据后,将对应的网络流量数据进行整理。

[0053] 步骤22、将网络流量数据转化预设的键值对key-value格式,得到网络流量时间序列,其中,key-value中key用于表征一个目标业务,key-value中value用于表征目标业务在预设时间长度中,每个单位时段接入的网络流量数据。

[0054] 在本实施例,在读取对应的网络流量数据后,将对应的网络流量数据进行整理,以将网络流量数据转换为key-value格式,其中,每个key代表了一个具体业务,key对应的value代表了这个业务在预设时间长度(例如:两个月)中,单位时段(例如:每个小时)接入的网络流量数据(也就是对应的流量值),然后按list形式存储网络流量时间序列。

[0055] 通过上述步骤中的从预设文件系统中获取对应的网络流量数据;将所述网络流量数据转化预设的键值对key-value格式,得到所述网络流量时间序列,实现了将待识别的网络流量数据的标准格式转换,形成时间序列的同时,满足处理系统的数据格式需求,提高异

常流量识别的效率。

[0056] 其中一些实施例中，步骤S102中的对网络流量时间序列进行预处理，可以通过如下步骤实现：

[0057] 步骤31、检测网络流量时间序列的缺失值。

[0058] 在本实施例中，缺失值是网络流量时间序列所对应的网络流量数据的缺失。

[0059] 在本实施例中，在业务新进接入时，对应的网络流量数据会造成一定的丢失，在数据写入过程中的某些时段，会造成写入数据丢失，因此，需要基于网络流量时间序列对应的网络流量数据的缺失值进行处理，可以理解的，需要在对缺失值进行处理之前，检测对应的网络流量时间序列的缺失值。

[0060] 步骤32、基于缺失值，对网络流量时间序列进行过滤处理，得到候选网络流量时间序列。

[0061] 在本实施例中，对应的过滤处理包括过滤掉缺失值大于设定阈值的网络流量时间序列、对写入过程中数据丢失造成的空值进行插值填充。

[0062] 在本实施例中，缺失值大于设定阈值的网络流量时间序列对应的网络流量数据可以认定为新进业务所接入的网络流量数据；将缺失值前后设定时长(例如：12小时)内的网络流量数据的均值作为插入值填充至空值中；在其中一些可选实施方式中，对于一段时间内写入的数据都出现问题的插值填充，则采用设定的默认值作为缺少值进行填充。

[0063] 步骤33、根据候选网络流量时间序列，生成多种类别的第一流量时序数据，其中，每种第一流量时序数据对应一种设定的变化周期。

[0064] 在本实施例中，在网络流量时间序列的数据本身存在的缺陷进行处理后，需要对网络流量时间序列对应的网络流量数据中会对异常流量预测造成影响的数据进行处理；在本实施例中，会将候选网络流量时间序列按变化趋势类别进行分解(例如：STL分解)，分解出不同趋势且变化周期不同的流量时序数据，也就是经过分解生成第一流量时序数据，以方便后续过滤会影响异常流量时间序列。

[0065] 步骤34、过滤多种第一流量时序数据中变化周期小于流量变化周期阈值的第一流量时序数据，并基于过滤得到的第二流量时序数据，生成网络流量时序数据集。

[0066] 在本实施例中，生成的多种第一流量时序数据中，包括长期时间趋势、季节性时间趋势、周期性时间趋势的时序数据和对应的残差项，而对应周期性时间趋势的时序数据，其对应的周期性中会存在变化周期较小的时序数据，例如：以天为变化周期的时序数据，该时序数据在进行后续处理后用于异常流量预测时，会提升误报率，因此，需要将对应的变化周期较小的时序数据进行过滤。

[0067] 在本实施例中，通过过滤掉变化周期较小的时序数据(例如：以天为变化周期的时序数据)，能使的在进行Z变换和时序数据增窗(X变换)后，输入到预测模型的时序数据在数据量级上保持一致，例如：部分业务在一天内的网络流量数据的流量值可以从接近于0上涨到10^5的数量级，在不作STL分解，并直接做增窗，一方面会掩盖掉一部分数量级较小的网络流量数据，另一方面极大的提升误报率，即当每日网络流量数据的峰值的数量级很大时(但为正常流量)，增窗下的该种网络流量数据会保持较大的绝对数值，从而被预测模型认为是异常流量；本实施例中，通过进行STL分解，可以避免异常值对于周期性规律估计的影响。

[0068] 通过上述步骤中的检测网络流量时间序列的缺失值;基于缺失值,对网络流量时间序列进行过滤处理,得到候选网络流量时间序列;根据候选网络流量时间序列,生成多种类别的第一流量时序数据,每种第一流量时序数据对应一种设定的变化周期;过滤多种第一流量时序数据中变化周期小于流量变化周期阈值的第一流量时序数据,并基于过滤得到的第二流量时序数据,生成网络流量时序数据集,实现了过滤变化周期小的时序数据的过滤,避免变化周期小的时序数据的周期性规律影响异常流量的预测,进一步降低异常流量识别的误报率。

[0069] 其中一些实施例中,步骤32中的基于缺失值,对网络流量时间序列进行过滤处理,得到候选网络流量时间序列,包括如下步骤:删除缺失值大于缺失值阈值的网络流量时间序列,得到候选网络流量时间序列,和/或,利用滑动平均算法对缺失值进行插值填充,得到候选网络流量时间序列。

[0070] 在本实施例中,缺失值大于设定阈值的网络流量时间序列对应的网络流量数据可以认定为新进业务所接入的网络流量数据,对于对应的网络流量数据而言,该部分网络流量数据可以认定为是异常的,通过删除对应的网络流量数据,减少无用预测次数,提高异常流量识别的效率。

[0071] 在本实施例中,滑动平均算法采用滑动/移动平均Moving Average,进一步可以优选简单移动平均SMA,并将缺失值前后设定时长(例如:12小时)内的网络流量数据的均值作为插入值填充至空值中;在其中一些可选实施方式中,对于一段时间内写入的数据都出现问题的插值填充,则采用设定的默认值作为缺少值进行填充。

[0072] 在本实施例中,通过基于缺失值进行过滤和缺失值插值,使的后续输入至预测网络中时序数据的完整性和有效性,提高异常流量识别的效率及降低误报率。

[0073] 其中一些实施例中,步骤33中的根据候选网络流量时间序列,生成多种类别的第一流量时序数据,可以通过如下步骤实现:利用时间序列分解算法STL对候选网络流量时间序列进行分解,生成多种第一流量时序数据。

[0074] 在本实施例中,通过时间序列分解算法STL将候选网络流量时间序列分解为长期时间趋势时序数据、季节性时间趋势时序数据、周期性时间趋势时序数据和残差项时序数据,在采用预测模型进行预测时,通过将代表异常的残差项时序数据输入对应的预测模型,则可以实现网络流量数据异常的快速预测,并通过生成多种类别的第一流量时序数据,可以降低目标时序数据的数据量维度,通过减少无意义的时序数据输入预测模型,并减少数据内存消耗和训练耗时,提高异常流量识别的处理效率。

[0075] 其中一些实施例中,步骤34中的基于过滤得到的第二流量时序数据,生成网络流量时序数据集,可以通过如下步骤实现:利用移动窗递推算法,对所有第二流量时序数据进行加窗,得到每条第二流量时序数据所对应的子时序数据集,其中,网络流量时序数据集包括每条第二流量时序数据所对应的子时序数据集。

[0076] 在本实施例中,在过滤得到第二流量时序数据后,会至少基于第二流量时序数据中残差项时序数据进行生成网络流量时序数据集的生成,也就是对残差项时序数据进行整理,将其转换为网络流量时序数据集,并用作异常流量预测的数据输入,当然,对残差项时序数据进行整理后,生成的网络流量时序数据集对应的也是残差项时序数据关联的时序数据。

[0077] 在本实施例中,采用固定长度的移动窗口(moving window with fixed length)对所有第二流量时序数据进行加窗。

[0078] 在其中一些可选实施方式中,设定一个第二流量时序数据对应的时间序列长L=8,窗口长度(wL)为3,则增窗后,可以获得5个时序数据,例如:设定对应的时序数据序列为[1,2,3,4,5,6,7,8],生成的5个时序数据构成的子时序数据集为

[0079] [[x:[1,2,3],y,[4]],

[0080] [[x:[2,3,4],y,[5]],

[0081] [[x:[3,4,5],y,[6]],

[0082] [[x:[4,5,6],y,[7]],

[0083] [[x:[5,6,7],y,[8]]。

[0084] 需要说明的是,本实施例中所解决的问题为在在时间序列分析中,无法通过单一一条网络流量数据预测及解析下一条网络流量数据异常,因此,通过设置移动窗口,通过这个移动窗口的所有网络流量数据来建模解释紧随这个窗口之后的网络流量数据是否异常,减少异常流量预测的误报率,提高预测的正确率。

[0085] 其中一些实施例中,第二流量时序数据至少包括残差项时序数据,步骤S102中的从预处理所生成的网络流量时序数据集中获取目标时序数据,包括如下步骤:

[0086] 步骤41、从网络流量时序数据集所对应的多种类别的第二流量时序数据中,选取残差项时序数据。

[0087] 步骤42、将残差项时序数据作为目标时序数据,其中,目标时序数据包括每条残差项时序数据对应的子时序数据集。

[0088] 通过上述步骤从网络流量时序数据集所对应的多种类别的第二流量时序数据中,选取残差项时序数据,将残差项时序数据作为目标时序数据,其中,目标时序数据包括每条残差项时序数据对应的子时序数据集,实现对目标时序数据的获取。

[0089] 其中一些实施例中,步骤S104中的基于预测时序数据和目标时序数据的误差值,确定异常流量识别结果,通过如下步骤实现:

[0090] 步骤51、确定残差项时序数据所对应多个误差值,其中,误差值为残差项时序数据与所对应的预测时序数据的差值的绝对值。

[0091] 在本实施例中,根据残差项时序数据对应的预测值与该残差项时序数据的实际值(预处理后输入到预测模型的残差项时序数据对应的值)的误差值,也就是预测值与实际值的差值的绝对值的平均值MAE,判断对应的网络流量数据是否为异常流量。

[0092] 步骤52、判断多个误差值之和是否大于预设误差值阈值。

[0093] 步骤53、在判断到多个误差值之和大于预设误差值阈值的情况下,确定网络流量时间序列所对应的网络流量数据为异常网络流量数据。

[0094] 通过上述步骤中的确定残差项时序数据所对应多个误差值,误差值为残差项时序数据与所对应的预测时序数据的差值的绝对值;判断多个误差值之和是否大于预设误差值阈值;在判断到多个误差值之和大于预设误差值阈值的情况下,确定网络流量时间序列所对应的网络流量数据为异常网络流量数,实现了异常流量的识别及预警。

[0095] 其中一些实施例中,长短期记忆神经网络预测模型包括基于注意力的长短期记忆网络Attention-based LSTM,长短期记忆神经网络预测模型的训练过程包括如下步骤:

[0096] 步骤61、获取历史网络流量数据，并将网络流量数据按键值对key-value格式转换为对应的历史网络流量时间序列。

[0097] 步骤62、对历史网络流量时间序列进行对应的预处理，得到历史网络流量时间序列所对应的残差项时间序列，其中，预处理包括缺失值过滤、插值、STL分解。

[0098] 在本实施例中，对于每项表征业务接入的网络流量的流量时间序列，STL分解可以初步过滤掉以天为周期的周期性，从而得到残差项时间序列，基于残差项时间序列的长短期记忆神经网络预测模型的训练难度会大幅度降低，且速度有明显提升，长短期记忆神经网络预测模型不需要再去学习以天为周期的周期性，并且对于残差项时间序列的正则化会更为有效，因为残差项时间序列的方差相比于原时间序列会大幅降低。

[0099] 在本实施例中，通过STL分解后过滤掉变化周期较小的时序数据(例如：以天为变化周期的时序数据)，能使的在Z变换和时序数据增窗(X变换)，并输入到预测模型的时序数据在数据量级上保持一致，例如：部分业务在一天内的网络流量数据的流量值可以从接近于0上涨到10^5的数量级，在不作STL分解，并直接做增窗，一方面会掩盖掉一部分数量级较小的网络流量数据，另一方面极大的提升误报率，即当每日网络流量数据的峰值的数量级很大时(但为正常流量)，增窗下的该种网络流量数据会保持较大的绝对数值，从而被预测模型认为是异常流量；本实施例中，通过进行STL分解，可以避免异常值对于周期性规律估计的影响。

[0100] 步骤63、对残差项时间序列进行增窗，得到对应的初始全量数据集，并按预设分配比将初始全量数据集拆解为训练集和测试集。

[0101] 在本实施例中，预设分配比可设定，例如：7:3。

[0102] 步骤64、通过训练集和测试集对初始Attention-based LSTM进行训练，得到初始全量数据集的权重和对应的待选预测模型。

[0103] 步骤65、基于初始全量数据集的权重，对初始全量数据集进行加权，得到备选全量数据集；

[0104] 步骤66、在备选全量数据集中，按权重由大到小顺序选取预设数据量的数据，得到目标全量数据，并通过目标全量数据对待选预测模型进行重训练，得到长短期记忆神经网络预测模型。

[0105] 在本实施例中，基于加权后的数据进行样本筛选，删除对应拟合较差的样本数据，减小预测模型基于错误的知识进行训练，从而造成预测模型预测的异常流量的误报率高。

[0106] 需要说明的是，在本实施例中，对历史网络流量时间序列的预处理过程可以参考本申请实施例中对采集的网络流量时间序列的预处理过程，此处不再过多赘述。

[0107] 需要说明的是，在本实施例中，对于每项业务的网络流量时间序列，STL分解可以初步过滤掉以天为周期的周期性，从而得到残差项时间序列；基于残差项时间序列的LSTM模型的训练难度会大幅度降低，且提升训练速度，并且对于残差项时间序列的正则化会更为有效，残差项时间序列的方差相比于原时间序列会大幅降低。同时，对于残差项时间序列中所蕴涵的较为复杂的周期性，使用Attention-based LSTM进行学习，用深度学习捕捉的规则更为贴合数据；在本实施例中，为了避免预测模型过度地从异常值中学习信息，对训练样本点进行加权，缓解从异常值学习异常知识的问题。

[0108] 图2是根据本申请优选实施例的训练Attention-based LSTM的流程示意图，如图2

所示，该流程包括如下步骤：

[0109] 步骤S201、从HDFS获取网络流量数据，之后，执行步骤S202。

[0110] 步骤S202、对每个业务对应的网络流量数据按每个小时统计，生成对应的网络流量时间序列，之后，执行步骤S203。

[0111] 步骤S203、基于网络流量时间序列所对应的缺失值，使用滑动平均进行插值，之后，执行步骤S204。

[0112] 步骤S204、对每项业务所对应的网络流量时间序列使用STL进行过滤，得到残差项时序数据，之后，执行步骤S205。

[0113] 步骤S205、对每项业务所对应的残差项时序数据使用Z变换，之后，执行步骤S206。

[0114] 步骤S206、对Z变换后的每项业务对应的残差时序数据进行增窗，得到Attention-based LSTM的总样本，之后，执行步骤S207。

[0115] 步骤S207、将总样本拆解为训练集与测试集，输入Attention-based LSTM，得到每个样本的权重，之后，执行步骤S208。

[0116] 步骤S208、根据每个样本的权中对总样本进行加权，并基于加权后的样本对Attention-based LSTM重训练。

[0117] 需要说明的是，在上述流程中或者附图的流程图中示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行，并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

[0118] 本实施例还提供了一种异常流量识别装置，该装置用于实现上述实施例及优选实施方式，已经进行过说明的不再赘述。如以下所使用的，术语“模块”、“单元”、“子单元”等可以实现预定功能的软件和/或硬件的组合。尽管以下实施例所描述的装置较佳地以软件来实现，但是硬件，或者软件和硬件的组合的实现也是可能并被构想的。

[0119] 图3是根据本申请实施例的一种异常流量识别装置的结构框图，如图3所示，该装置包括：

[0120] 获取模块31，用于获取网络流量时间序列，其中，网络流量时间序列用于表征目标业务在预设时间长度中，每个单位时段所接入的网络流量数据；

[0121] 预处理模块32，与获取模块31耦合连接，用于对网络流量时间序列进行预处理，并从预处理所生成的网络流量时序数据集中获取目标时序数据，其中，目标时序数据用于表征网络流量时间序列的异常数据；

[0122] 处理模块33，与预处理模块32耦合连接，用于利用长短期记忆神经网络预测模型对目标时序数据进行处理，获得与目标时序数据对应的预测时序数据，其中，长短期记忆神经网络预测模型被训练为用于根据输入的时序数据得到与该时序数据对应的预测数据；

[0123] 确定模块34，与处理模块33耦合连接，用于基于预测时序数据和目标时序数据的误差值，确定异常流量识别结果。

[0124] 通过本申请实施例的异常流量识别装置，采用获取网络流量时间序列，网络流量时间序列用于表征目标业务在预设时间长度中，每个单位时段所接入的网络流量数据；对网络流量时间序列进行预处理，并从预处理所生成的网络流量时序数据集中获取目标时序数据，目标时序数据用于表征网络流量时间序列的异常数据；利用长短期记忆神经网络预测模型对目标时序数据进行处理，获得与目标时序数据对应的预测时序数据；基于预测时

序数据和目标时序数据的误差值，确定异常流量识别结果，解决相关技术中异常流量识别困难、误报率高的问题，实现了降低业务流量异常监测误报率、减少因异常流量误报所造成的客诉及降低人力成本的有益效果。

[0125] 在其中一些实施例中，该获取模块31还用于：从预设文件系统中获取对应的网络流量数据；将网络流量数据转化预设的键值对key-value格式，得到网络流量时间序列，其中，key-value中key用于表征一个目标业务，key-value中value用于表征目标业务在预设时间长度中，每个单位时段接入的网络流量数据。

[0126] 在其中一些实施例中，该预处理模块32还进一步包括：

[0127] 第一检测单元，用于检测网络流量时间序列的缺失值；

[0128] 第一过滤单元，与第一检测单元耦合连接，用于基于缺失值，对网络流量时间序列进行过滤处理，得到候选网络流量时间序列；

[0129] 第一生成单元，与第一过滤单元耦合连接，用于根据候选网络流量时间序列，生成多种类别的第一流量时序数据，其中，每种第一流量时序数据对应一种设定的变化周期；

[0130] 第一处理单元，与第一生成单元耦合连接，用于过滤多种第一流量时序数据中变化周期小于流量变化周期阈值的第一流量时序数据，并基于过滤得到的第二流量时序数据，生成网络流量时序数据集。

[0131] 在其中一些实施例中，第一生成单元还用于利用时间序列分解算法STL对候选网络流量时间序列进行分解，生成多种第一流量时序数据。

[0132] 在其中一些实施例中，第一处理单元还用于利用移动窗递推算法，对所有第二流量时序数据进行加窗，得到每条第二流量时序数据所对应的子时序数据集，其中，网络流量时序数据集包括每条第二流量时序数据所对应的子时序数据集。

[0133] 在其中一些实施例中，第二流量时序数据至少包括残差项时序数据，该预处理模块32还用于从网络流量时序数据集所对应的多种类别的第二流量时序数据中，选取残差项时序数据；将残差项时序数据作为目标时序数据，其中，目标时序数据包括每条残差项时序数据对应的子时序数据集。

[0134] 在其中一些实施例中，该第一过滤单元还用于删除缺失值大于缺失值阈值的网络流量时间序列，得到候选网络流量时间序列，和/或，利用滑动平均算法对缺失值进行插值填充，得到候选网络流量时间序列。

[0135] 在其中一些实施例中，该确定模块34还进一步包括：

[0136] 第一确定单元，用于确定残差项时序数据所对应多个误差值，其中，误差值为残差项时序数据与所对应的预测时序数据的差值的绝对值；

[0137] 第一判断单元，与第一确定单元耦合连接，用于判断多个误差值之和是否大于预设误差值阈值；

[0138] 第一识别单元，与第一判断单元耦合连接，用于在判断到多个误差值之和大于预设误差值阈值的情况下，确定网络流量时间序列所对应的网络流量数据为异常网络流量数据。

[0139] 图4是本申请实施例的电子设备的结构示意图，如图4所示，本申请实施例提供了一种电子设备，包括处理器41、通信接口42、存储器43和通信总线44，其中，处理器41，通信接口42，存储器43通过通信总线44完成相互间的通信，

[0140] 存储器43,用于存放计算机程序;

[0141] 处理器41,用于执行存储器43上所存放的程序时,实现图1中的方法步骤。

[0142] 该电子设备中的处理实现图1中的方法步骤,所带来的技术效果与上述实施例执行图1中异常流量识别方法的技术效果一致,在此不再赘述。

[0143] 上述电子设备提到的通信总线可以是外设部件互连标准(Peripheral Component Interconnect,简称PCI)总线或扩展工业标准结构(Extended Industry Standard Architecture,简称EISA)总线等。该通信总线可以分为地址总线、数据总线、控制总线等。为便于表示,图4中仅用一条粗线表示,但并不表示仅有一根总线或一种类型的总线。

[0144] 通信接口用于上述终端与其他设备之间的通信。

[0145] 存储器可以包括随机存取存储器(RandomAccess Memory,简称RAM),也可以包括非易失性存储器(non-volatile memory),例如至少一个磁盘存储器。可选的,存储器还可以是至少一个位于远离前述处理器的存储装置。

[0146] 上述的处理器可以是通用处理器,包括中央处理器(Central Processing Unit,简称CPU)、网络处理器(Network Processor,简称NP)等;还可以是数字信号处理器(Digital Signal Processing,简称DSP)、专用集成电路(Application Specific Integrated Circuit,简称ASIC)、现场可编程门阵列(Field-Programmable Gate Array,简称FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。

[0147] 本申请实施例还提供了一种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前述任意一个方法实施例提供的异常流量识别方法的步骤。

[0148] 在本申请提供的又一实施例中,还提供了一种包含指令的计算机程序产品,当其在计算机上运行时,使得计算机执行上述实施例中任一所述的异常流量识别方法的步骤。

[0149] 需要说明的是,在本文中,诸如“第一”和“第二”等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0150] 以上所述仅是本发明的具体实施方式,使本领域技术人员能够理解或实现本发明。对这些实施例的多种修改对本领域的技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下,在其它实施例中实现。因此,本发明将不会被限制于本文所示的这些实施例,而是要符合与本文所申请的原理和新颖特点相一致的最宽的范围。

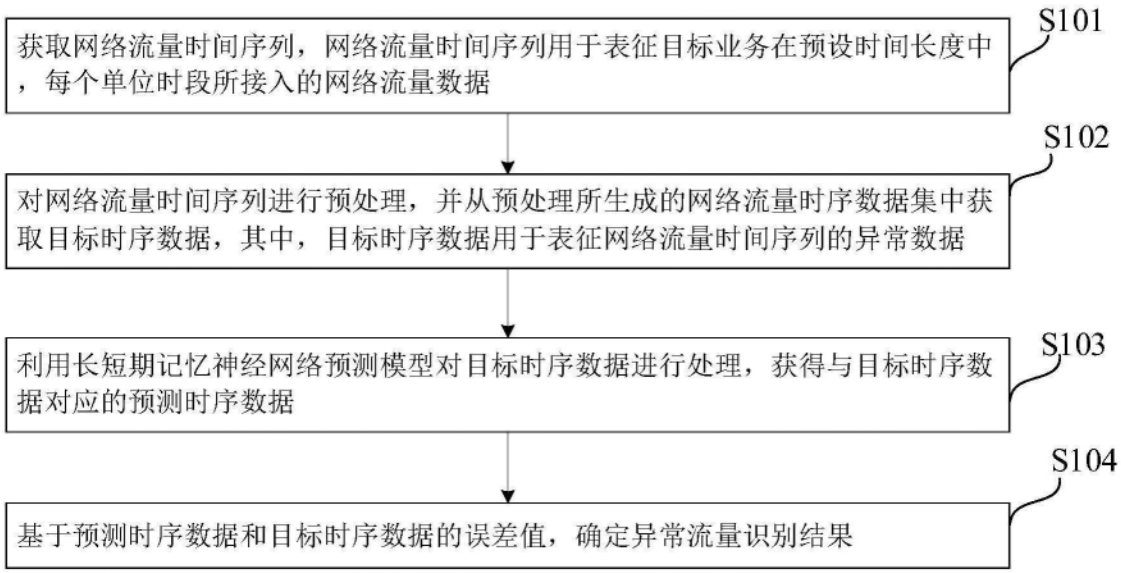

S101
获取网络流量时间序列，网络流量时间序列用于表征目标业务在预设时间长度中，每个单位时段所接入的网络流量数据
S102
对网络流量时间序列进行预处理，并从预处理所生成的网络流量时序数据集中获取目标时序数据，其中，目标时序数据用于表征网络流量时间序列的异常数据
S103
利用长短期记忆神经网络预测模型对目标时序数据进行处理，获得与目标时序数据对应的预测时序数据
S104
基于预测时序数据和目标时序数据的误差值，确定异常流量识别结果

图1

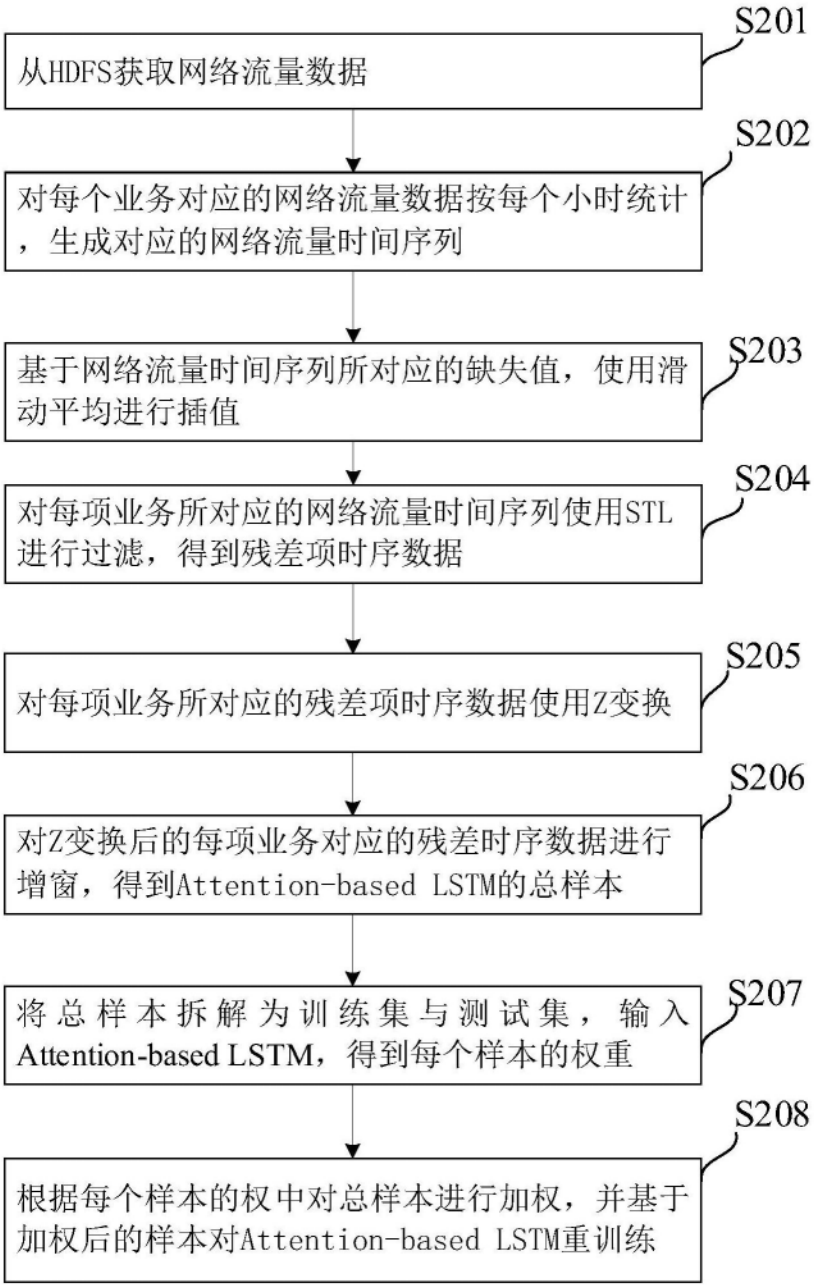

S201
从HDFS获取网络流量数据
S202
对每个业务对应的网络流量数据按每个小时统计，生成对应的网络流量时间序列
S203
基于网络流量时间序列所对应的缺失值，使用滑动平均进行插值
S204
对每项业务所对应的网络流量时间序列使用STL进行过滤，得到残差项时序数据
S205
对每项业务所对应的残差项时序数据使用Z变换
S206
对Z变换后的每项业务对应的残差时序数据进行增窗，得到Attention-based LSTM的总样本
S207
将总样本拆解为训练集与测试集，输入Attention-based LSTM，得到每个样本的权重
S208
根据每个样本的权中对总样本进行加权，并基于加权后的样本对Attention-based LSTM重训练

图2

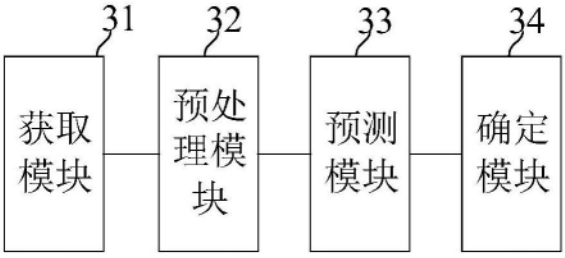

31
32
33
34
获取模块
预处理模块
预测模块
确定模块

图3

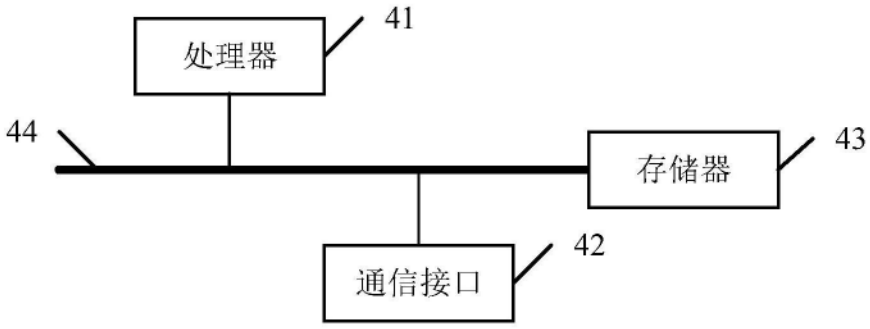


图4