Journal of Computer Applications
计算机应用,2020,40(5):1335 - 1339
ISSN 1001-9081
CODEN JYIIDU
2020-05-10
http://www.joca.cn

文章编号:1001-9081(2020)05-1335-05
DOI:10.11772/j.issn.1001-9081.2019111970

基于长短期记忆网络和滑动窗口的流数据异常检测方法

仇　媛[1],常相茂[1*],仇　倩[2],彭　程[1],苏善婷[1]
(1.南京航空航天大学 计算机科学与技术学院,南京 211106;　2.河北工业大学 人工智能与数据科学学院,天津 300401)
(*通信作者电子邮箱xiangmaoch@nuaa.edu.cn)

摘　要:针对目前流数据存在数量巨大、生成迅速和概念漂移的特点,提出了一种基于长短期记忆(LSTM)网络和滑动窗口的流数据异常检测方法。首先采用LSTM网络进行数据预测,之后计算预测值与实际值的差值。对于每个数据,选择合适的滑动窗口,将滑动窗口区间内的所有差值进行分布建模,再根据每个差值在当前分布的概率密度来计算数据异常可能性。LSTM网络不仅可以进行数据预测,还可以边预测边学习,实时更新调整网络,保证模型的有效性;而利用滑动窗口可以使得异常分数的分配更为合理。最后使用在真实数据基础上制造的模拟数据进行了实验。实验结果验证了所提方法在低噪声环境下比直接利用差值进行检测和异常数据分布建模法(ADM)方法的平均曲线下面积(AUC)值分别提高了0.187和0.05。

关键词:流数据;异常检测;滑动窗口;长短期记忆网络;神经网络
中图分类号:TP391.4　　**文献标志码**:A

Stream data anomaly detection method based on long short-term memory network and sliding window

QIU Yuan[1], Chang Xiangmao[1*], QIU Qian[2], PENG Cheng[1], SU Shanting[1]
(1. *College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing Jiangsu* 211106, *China*;
2. *College of Artificial Intelligence and Data Science, Hebei University of Technology, Tianjin* 300401, *China*)

Abstract: Aiming at the characteristics of large volume, rapid generation and concept drift of current stream data, a stream data anomaly detection method based on Long Short-Term Memory (LSTM) network and sliding window was proposed. Firstly, the LSTM network was used for data prediction, and the difference between the predicted value and the actual value was calculated. For each datum, the appropriate sliding window was selected, and the distribution modeling was performed to all the differences in the sliding window interval, then the probability of data anomaly was calculated according to the probability density of each difference in the current distribution. The LSTM network was not only able to predict data, but also able to predict and learn at the same time, as well as update and adjust the network in real time to ensure the validity of the model. The use of sliding windows was able to make the allocation of abnormal scores more reasonable. Finally, the simulation data made on the basis of real data were used for experiment. The experimental results verify that the average Area Under Curve (AUC) value of the proposed method in low-noise environment is 0.187 and 0.05 higher than that of direct difference detection and Abnormal data Distribution Modeling (ADM) method, respectively.

Key words: stream data; anomaly detection; sliding window; Long Short-Term Memory (LSTM) network; neural network

0　引言

随着传感器、监视测量技术的广泛应用和网络交易、安全监控、电信管理等应用的不断发展,产生了大量高速实时数据的无界序列,称为数据流。这些数据流中的数据被称为流数据[1-2]。在对流数据的分析处理中,异常检测一直是被广大研究人员重点关注的领域,特别是工业设备的监控、军事项目管理或是网络安全监测等,如果异常没有被及时发现,更是可能会造成破坏性的后果[3]。现阶段,流数据的大量增加也为数据的异常检测提出了新的要求。

流数据通常数量巨大且生成迅速,因此要求针对它的异常检测方法必须是实时的;其次,系统的统计数据可能会随着时间的推移而发生变化,这种现象被称为概念漂移[1]。传统的利用正常数据训练完成的模型进行异常检测的方法不再适用,模型必须做到及时更新来保证模型的检测有效。本文提出了一种基于长短期记忆(Long Short-Term Memory, LSTM)网络的异常检测方法,通过LSTM网络进行数据预测和异常识别,并利用滑动窗口对预测差值进行分布建模,为每个数据分配更加精准的异常分数,从而实现更加精准的异常检测。

异常是与历史模式不一致或偏离预期的一般比较罕见的事件,以至于怀疑它是由与正常数据不同的机制产生的,而异

收稿日期:2019-11-04;**修回日期**:2019-11-25;**录用日期**:2019-11-26。
作者简介:仇媛(1995—),女,河北石家庄人,硕士研究生,CCF会员,主要研究方向:异常检测、深度学习;　常相茂(1982—),男,山东淄博人,副教授,博士,主要研究方向:物联网、基于可穿戴设备的智能健康监测、机器学习算法的感知数据处理及分析;　仇倩(1995—),女,河北石家庄人,硕士研究生,主要研究方向:社会网络、深度学习;　彭程(1995—),男,安徽合肥人,硕士研究生,CCF会员,主要研究方向:状态检测、深度学习;　苏善婷(1994—),女,江苏常熟人,硕士研究生,CCF会员,主要研究方向:故障检测、机器学习。

常值检测就被定义为“发现”这些与其他观察结果偏离的观察结果。到目前为止，异常检测的方法有很多。根据检测类别的不同可以将异常检测分为异常序列检测、异常子序列检测和单个异常点检测[4]，因为想要找出流数据中的异常点，且为每个数据分配一个合理的异常分数来展示它是异常的可能性，所以本文要实现的是单个异常点的检测。此外，根据检测方法的不同[5]，异常检测还可以分为基于统计的检测方法[6-7]、基于距离的检测方法[8-9]、基于密度的检测方法[10-11]、基于聚类的检测方法[12-13]、基于分类的检测方法[14-15]等。这些方法各自有各自的优势和适用领域，但正如之前所说，绝大多数异常检测算法适用于批量处理数据，而由于流数据生成迅速、数量庞大和概念漂移的特点，这些传统的异常检测方法不能直接应用于流数据。

当然，近几年也提出了一些针对流数据的异常检测方法。一些在线异常检测方法，如大多数聚类算法、分布式分组匹配算法[16]、数据流多类学习算法[17]等，虽然可以满足流数据异常检测所需的实时性，但它们无法解决流数据的概念漂移问题。此外，Yu等[18]根据网络流量模式变化的特点提出了针对于交通模式变化的异常检测方法，该算法基于半监督技术，利用数据流模型训练检测模型来进行检测。这属于针对特定领域开发的基于模型的方法，需要明确的领域知识，适用性不高，不易推广。卡尔曼滤波技术也属于特定领域的方法，因为它需要相应领域的知识来调整参数并选择对应的残差模型[19]。除此之外，还有一些轻量级的统计方法，比如指数平滑[20]、变换点检测[21]等，但这些方法大都集中检测空间异常，在具有时间依赖性的应用中可用性不高，而针对时间序列建模的整合移动平均自回归模型（Autoregressive Integrated Moving Average model，ARIMA）[22]是用季节性来进行时态数据建模，适合常规的每日或每周模式检测数据中的异常，但如果数据没有这种季节周期性就不再适用。

基于此，本文提出了一种基于LSTM网络和滑动窗口的异常检测方法。LSTM是神经网络的一种，可以进行数据预测，相较于其他网络，LSTM可以快速学习数据的新特征，实时更新调整网络，保证模型的有效性，避免因流数据概念漂移问题导致的神经网络预测不准确。此外，选取一个滑动窗口，将窗口内所有差值进行分布建模来为每个数据分配异常分数，可以直观地观察每个数据是异常的可能性大小。

1 模型构建

1.1 方法模型

用x_t表示t时刻模型接收到的实时数据，x_t可能是来自于传感器网路的传感器数据、各种服务器的CPU使用率、零售业交易数据、带宽的测量值等，因此，模型的输入为$x_t, x_{t+1}, x_{t+2}, \cdots$。

本文所要解决的问题就是如何检测流数据中的异常数据，以及如何使数据分析人员可以更加直观地观察出每个数据的异常可能性。为了定义该异常可能性，本方法为每一个输入数据x_t分配异常分数$AS_t(AS_t \in [0,1])$，AS_t越大，x_t为异常数据的可能性就越大。为了标记异常数据，为每一个输入数据x_t分配异常属性值a_t，将得到的异常分数AS_t跟阈值T比较：大于阈值则x_t为异常数据，$a_t = 1$；否则，x_t为正常数据，$a_t = 0$。a_t为x_t是否为异常的判断输出，仅有0或1两个取值。

1.2 LSTM网络

LSTM是神经网络的一种，神经网络是一种通过自身相互连接的神经元们来进行函数近似，从而进行机器学习或模式识别的自适应系统。它可以通过输入大量训练数据进行多次迭代来调整自身参数，得到一个学习到训练数据特征的模型。当有新数据输入该模型时，模型便可根据参数计算相应的输出，完成分类或是预测任务。神经网络通常由输入层、隐含层和输出层组成，参数调整指的便是各个层之间神经元的连接函数中参数的改变。

传统的神经网络，如循环神经网络（Recurrent Neural Network，RNN），隐含层只有一个状态，对短期的输入十分敏感，存在梯度消失的问题，不适合处理长序列数据，特别是时间序列数据。LSTM网络在隐含层中层加了一个状态来存储长期状态，避免了RNN的梯度消失。因此，LSTM特别适合处理时间序列数据，广泛应用于时间序列数据的预测和异常检测[23]。

LSTM单元的逻辑架构示意图如图1所示。当t时刻的数据x_t输入LSTM时，便与$t-1$时刻的长期状态$\boldsymbol{c}_{t-1}$和输出$\boldsymbol{h}_{t-1}$一起作为输入参与运算，得到t时刻LSTM的输出$\boldsymbol{h}_t$，$\boldsymbol{h}_t$经过维度转化后便为所求的t时刻的预测数据。

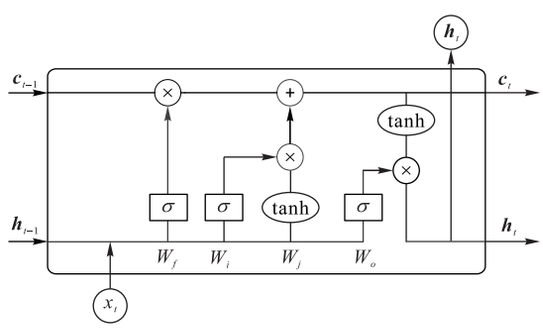

图1 LSTM单元逻辑架构示意图

Fig. 1 Schematic diagram of LSTM unit logical architecture

LSTM的关键之处就在于控制长期状态$\boldsymbol{c}_t$，令网络的每一个输出都是长期状态参与运算之后的结果，避免梯度消失问题。为此，LSTM设置了3个门：遗忘门、输入门和输出门。先是遗忘门，它用来控制LSTM是否以一定概率遗忘上一层隐藏细胞状态。其次是输入门，它用来处理当前序列位置的输入，当前时刻的单元状态$\boldsymbol{c}_t$在该阶段得到更新，$\boldsymbol{c}_t$的计算公式如式(1)所示：

$$\boldsymbol{c}_t = \text{sigmod}(\boldsymbol{f}_t) \times \boldsymbol{c}_{t-1} + \text{sigmod}(\boldsymbol{i}_t) \times \tanh(\boldsymbol{c}_t) \tag{1}$$

经过这个阶段，LSTM当前记忆与长期记忆组合完成了单元状态$\boldsymbol{c}_t$的更新。最后是输出门，它用来计算当前序列位置的输出，在该阶段，计算当前时刻的输出$\boldsymbol{h}_t$，$\boldsymbol{h}_t$的计算公式如式(2)所示：

$$\boldsymbol{h}_t = \text{sigmod}(\boldsymbol{o}_t) \times \tanh(\boldsymbol{c}_t) \tag{2}$$

输出门和单元状态$\boldsymbol{c}_t$共同确定了LSTM的输出。

图1仅是LSTM单元的逻辑架构图，实际上每处运算均由λ个神经元参与（λ为隐含层神经元的数量），LSTM的输出也为λ维向量，因此，LSTM模型需要最后连接一个全连接层，将LSTM层的输出进行维度变化，使结果呈现想要的维度。

与其他神经网络相同，堆叠隐含层可以使模型更加深入，从而得到更加准确的输出。多个LSTM层组成的LSTM模型称为堆叠式LSTM模型，其模型结构如图2所示。

图2　堆叠式LSTM模型结构

Fig. 2　Stacked LSTM model structure

2　方法设计

为了解决1.1节中提出的问题,本文提出了基于LSTM和滑动窗口的流数据异常检测方法SDLS(Stream data anomaly Detection method based on LSTM and Sliding window)。运用LSTM网络进行数据预测,且由于LSTM单元具有遗忘门,网络可以快速学习数据的新特征,实时更新调整网络,避免因流数据概念漂移问题导致的神经网络预测不准确。此外,对于每一个数据,本方法不直接将预测差值作为其异常分数,而是选取一个合适大小的滑动窗口,将窗口内所有差值进行分布建模,再根据差值在分布内的概率密度值来分配当前数据的异常分数大小,解决流数据概念漂移问题导致的异常分数匹配不合理现象。图3显示了SDLS的架构。

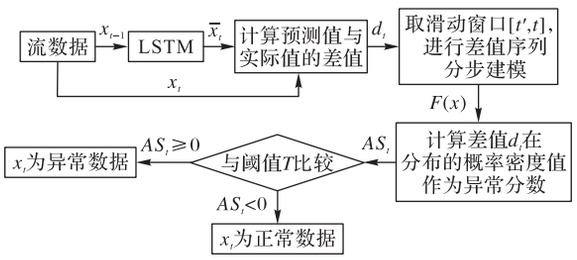

图3　SDLS方法整体架构示意图

Fig. 3　Schematic diagram of overall architecture of SDLS method

2.1　预测差值计算

本文选择堆叠式LSTM网络进行数据预测。LSTM的预测过程如下所示:

1)x_t作为输入进入第一个LSTM层;

2)x_t与$\boldsymbol{h}_{t-1,1}$进行连接,得到$[x_t, \boldsymbol{h}_{t-1,1}]$,$[x_t, \boldsymbol{h}_{t-1,1}]$与矩阵$\boldsymbol{W}(f_1)$进行矩阵点乘,得到结果矩阵$\boldsymbol{f}_{t,1}$,$\boldsymbol{f}_{t,1}$再进行sigmod操作,得到$\mathrm{sigmod}(\boldsymbol{f}_{t,1})$。

3)$[x_t, \boldsymbol{h}_{t-1,1}]$与矩阵$\boldsymbol{W}(\boldsymbol{i}_1)$进行矩阵点乘,得到结果矩阵$\boldsymbol{i}_{t,1}$,$\boldsymbol{i}_{t,1}$再进行sigmod操作,得到$\mathrm{sigmod}(\boldsymbol{i}_{t,1})$;$[x_t, \boldsymbol{h}_{t-1,1}]$与矩阵$\boldsymbol{W}(\boldsymbol{j}_1)$进行矩阵点乘,得到结果矩阵$\boldsymbol{j}_{t,1}$,$\boldsymbol{j}_{t,1}$再进行tanh操作,得到$\tanh(\boldsymbol{j}_{t,1})$;最后根据式(1)计算当前时刻的单元状态$\boldsymbol{c}_{t,1}$。

4)$[x_t, \boldsymbol{h}_{t-1,1}]$与矩阵$\boldsymbol{W}(\boldsymbol{o}_1)$进行矩阵点乘,得到结果矩阵$\boldsymbol{o}_{t,1}$,$\boldsymbol{o}_{t,1}$再进行sigmod操作,得到$\mathrm{sigmod}(\boldsymbol{o}_{t,1})$。最后根据式(2)计算当前时刻的输出$\boldsymbol{h}_{t,1}$。

5)$\boldsymbol{h}_{t,1}$作为输入传入第下一个LSTM层,重复进行步骤2)~4),直到得到$\boldsymbol{h}_{t,m}$(m为LSTM隐含层的层数)。

6)$\boldsymbol{h}_{t,m}$输入全连接层,进行维度转换,得到t+1时刻的预测输出$\bar{x}_{t+1}$。

2.2　异常分数计算

通常情况下,计算预测值$\bar{x}_t$与实际值x_t的差值d_t,将差值d_t转化为统一度量的异常分数$AS_t(AS_t \in [0,1])$,d_t越大,异常分数AS_t越大,然后将异常分数与规定的阈值T比较,大于T的为异常数据。但是,由于流数据的概念漂移现象,正常数据和异常数据的模式都会发生变化,单纯地将以差值的大小来计算异常分数会使得发生概念漂移之后的异常分数不准确。因此,选取一个合适大小的滑动窗口,将滑动窗口区间内差值序列建模为正态分布,再计算差值d_t在该分布内的概率密度值作为x_t的异常分数AS_t。这样,x_t的异常分数取决于在t时刻及其之前一段时间内的预测差值分布,异常分数的分配可以随着流数据的变化而变化,准确性更高。

对于每一个数据x_t,计算异常分数首先要为其选取合适大小的滑动窗口。因为异常数据的出现通常代表数据出现与之前不一致的模式,本方法选取倒数第n个异常所在时刻t_{an}到t时刻的区间作为建模分布的区间,即$[t_{an}, t]$。维持这个区间可以令滑动窗口区间内差值序列的分布建模更加符合流数据的实时数据模式,从而得到更加适合的分布函数。然而,如果某段时间内出现连续不断的异常,那么最后n个异常所在区间会非常小,用该区间进行建模也会令得到的分布函数因数据量过小而准确率不高,因此,本方法规定了滑动窗口的最小长度l,如果$t_{an} \geqslant t-l$,则选取$[t-l, t]$作为滑动窗口的区间范围。滑动窗口的选取有两个标准:第一个标准是t时刻之前最后n个异常所在的区间$[t_{an}, t]$;第二个标准为t时刻之前固定长度的区间$[t-l, t]$,$t' = \min\{t_{an}, t-l\}$,$[t',\ t]$作为最终的滑动窗口区间。

举例来说明该过程。图4为一个时间序列异常分数的分布示意图,令阈值T为0.5,则AS值在虚线以上的数据为异常数据。令t=50,n=10,l=30,则按照第一个标准选出的滑动窗口区间是50 s之前最后10个异常所在的区间,即为图中S_2区间;按照第二个标准选出的滑动窗口区间是20~50 s区间,即图中S_1区间,因为20 s<30 s,本文取[20 s,50 s]作为最终的滑动窗口区间。

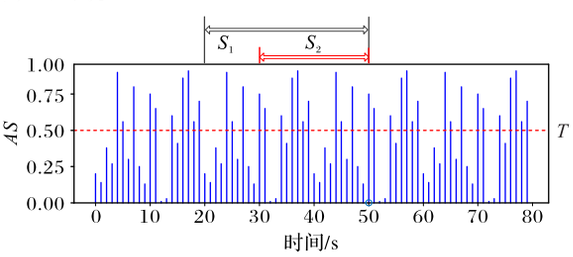

图4　计算t=50时数据的异常分数时按照两个标准选取的滑动区间

Fig. 4　Sliding intervals selected according to two criterias when calculating abnormal score of data at time t=50

选择好滑动窗口后,计算$[t', t]$区间内差值序列$[d_{t'}, d_t]$的平均值μ和方差σ^2,再用μ与σ^2进行正态分布建模,得到$F(x) = N\sim(\mu, \sigma^2)$,再计算差值$d_t$在该分布内的概率密度值$F(d_t)$,$F(d_t)$便为$x_t$的异常分数$AS_t$。

确定完异常分数后,将异常分数与阈值T进行比较,大于阈值是异常数据,否则为正常数据。由此,完成了异常数据的识别。

综上,异常数据检测算法的伪代码如下所示。

算法1　异常数据检测。

输入　流数据序列$x_t, x_{t+1}, \cdots$;

输出　流数据异常判定序列:$a_t, a_{t+1}, \cdots, a_i \in \{0,1\}$(0为正常,1为异常),及流数据异常分数序列$AS_t, AS_{t+1}, \cdots$。

1)将流数据序列$x_t, x_{t+1}, \cdots$输入LSTM,得到预测数据$\bar{x}_t, \bar{x}_{t+1}, \cdots$

2)计算预测差值,$d_i = |\bar{x}_i - x_i|, i = t, t+1, \cdots$

3)记录最初的n个异常值所在时刻:$h_0, h_1, \cdots, h_{n-1}$,此时直接将

预测差值d转化为异常分数AS再与阈值比较来判断是否为异常值。

4) for i++:

d=[h_0, i], μ=mean(d), σ=standard(d)

F=N~(μ,σ^2), $AS_i = F(d_i)$

if $AS_i > T$: $a_i = 1$

else $a_i = 0$

if $s_i = 1$ & $i - h_1 > l$:

for k in 1,2,…,n−1:

$h_k = h_k + 1$

$h_n = i$

/*n为窗口的异常数量的大小,l为窗口最小长度*/

3 实验验证

本文使用液压系统状态检测数据集[24]中的压力传感器数据来评估本文方法的性能。该数据集是用液压实验台实验获得的,实验台由一个主要工作和一个二次冷却过滤回路组成,它们通过油箱连接。系统每60 s一个循环,同时测量循环过程中的压力、温度等数据值,在系统运行过程中,4个液压部件(冷却器、阀门、泵和蓄能器)的状态会发生改变,数据的走势也会随之发生改变。

首先将数据输入LSTM来得到预测数据,本实验所用LSTM含有3个隐含层,其中的神经元个数分别为128、64、16,时间步长设置为150。传感器数据通常包含三种噪声[25],本文在原始数据中添加这三种类型噪声制造含噪声数据,之后将含噪声数据输入LSTM,并改变噪声百分比来评估LSTM的预测精度,本文使用式(3)来评估训练的SDAE的拟合精度,Acu越大说明准确率越高。

$$Acu = 1 - \frac{1}{L}\sum_{t=1}^{L}\left|\frac{x_t - \bar{x}_t}{x_t}\right| \tag{3}$$

其中:x_t为原始数据,$\bar{x}_t$是LSTM通过含噪声数据预测的数据,L为时间序列长度。通过图5可以看到,虽然随着噪声百分比的增大,Acu也随之下降变大,但Acu始终在90%以上,说明LSTM预测拟合数据的能力很好。

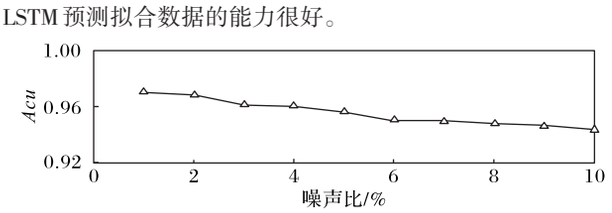

图5 LSTM在不同噪声百分比下预测拟合数据的Acu

Fig. 5 Acu of fitted data predicted by LSTM at different noise percentages

通过图6,可以直观地看到LSTM的预测能力,它可以很好地预测数据且可以很大程度上避免噪声的影响,因此,它的预测差值可以用来检测异常值。

在实验中,设置l=80,n=20,T=0.5,若$AS_t \geqslant 0.5$则说明x_t为异常值。

为了评估检测SDLS检测异常数据、确定异常分数的能力,本文将SDLS与直接差值异常检测法(Predicted Difference, PD)和另一种应用于流数据的异常检测方法——异常数据分布建模法(Abnormal Data distribution Modeling, ADM)[19]进行比较,PD直接将预测差值的大小转化为异常分数,预测差值越大,则异常分数越大。ADM为另一种用滑动窗口内的预测差值进行分布建模计算异常分数的方法,但ADM进行分布建模时仅考虑窗口内那些已经被判定为异常的数据,并且使用Q函数来计算异常分数。本文使用曲线下面积(Area Under Curve, AUC)来评估算法的性能,这是一种广泛使用的度量标准,用于评估离群值检测的性能[26]。当AUC较大时,算法的性能更好。

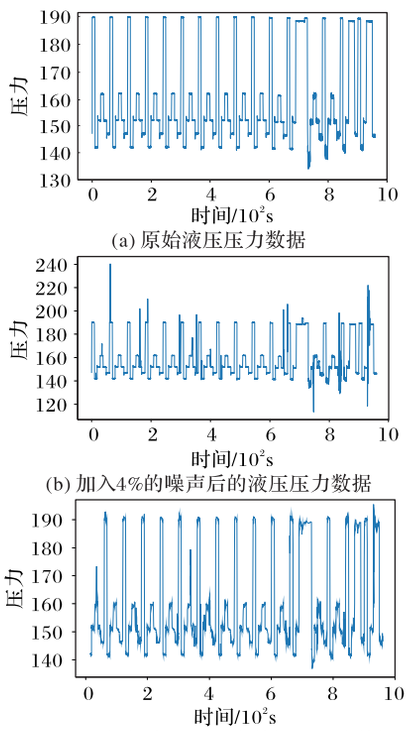

图6 LSTM对含噪声数据进行预测的效果

Fig. 6 Effect of LSTM predicting noisy data

图7显示了三种方法在应用于具有不同噪声百分比的数据时检测异常数据的性能比较,可以看到,SDLS一直优于其他两种方法。SDLS的平均AUC值与PD和ADM相比分别提高了0.187和0.05(SDLS方法的平均值为0.915 527,PD方法的平均值为0.865 537,ADM方法的平均值为0.728 742)。ADM表现不佳,因为ADM仅对异常数据进行分布建模,适用于有大量嘈杂的异常数据的应用,因此在异常数据较少时,AUC值不高。

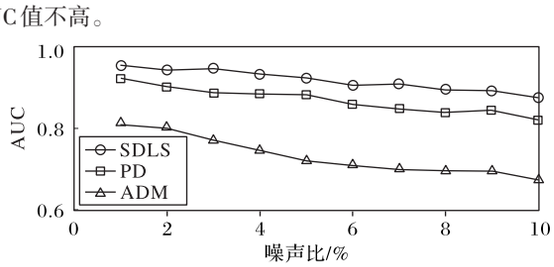

图7 三种流数据异常检测方法的AUC性能对比

Fig. 7 AUC performance comparison of three stream data anomaly detection methods

4 结语

针对流数据存在概念漂移的问题,本文提出了一种基于LSTM和滑动窗口的流数据异常检测方法SDLS。首先通过LSTM进行数据预测求出预测差值,再将滑动窗口内的差值序列进行分布建模,动态地为每个数据分配更加合适的异常分

数,提高流数据异常检测的准确率。本文利用真实实验数据构造含有噪声的测试数据,利用基于测试数据的实验验证了该方法的有效性。

参考文献 (References)

[1] WU K, ZHANG K, FAN W, et al. RS-Forest: a rapid density estimator for streaming anomaly detection[C]// Proceedings of the 2014 IEEE International Conference on Data Mining. Piscataway: IEEE, 2014:600-609.

[2] SUN D, HU Y, SHI Z, et al. An efficient anomaly detection framework for electromagnetic streaming data[C]// Proceedings of the 4th International Conference on Big Data and Computing. New York: ACM, 2019:151-155.

[3] DING Z, FEI M. An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window[J]. IFAC Proceedings Volumes, 2013, 46(20):12-17.

[4] GUPTA M, GAO J, AGGARWAL C C, et al. Outlier detection for temporal data: a survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(9):2250-2267.

[5] THAKKAR P, VALA J, PRAJAPATI V. Survey on outlier detection in data stream[J]. International Journal of Computer Applications, 2016, 136(2):13-16.

[6] LI X, HAN J. Mining approximate top-k subspace anomalies in multi-dimensional time-series data[EB/OL]. [2019-05-20]. https://pdfs. semanticscholar. org/fe68/274342529b66baa0f38e026b607b63faba9c. pdf.

[7] SOLBERG H E, LAHTI A. Detection of outliers in reference distributions: performance of Horn's algorithm[J]. Clinical Chemistry, 2005, 51(12):2326-2332.

[8] 李春生,于澍,刘小刚. 基于改进距离和的异常点检测算法研究[J]. 计算机技术与发展, 2019, 29(3):97-100. (LI C S, YU S, LIU X G. Research on outlier detection algorithm based on improved distance [J]. Computer Technology and Development, 2019, 29(3):97-100.)

[9] 蒋华,张红福,罗一迪,等. 基于KL距离的自适应阈值网络流量异常检测[J]. 计算机工程, 2019, 45(4):108-113, 118. (JIANG H, ZHANG H F, LUO Y D, et al. Adaptive threshold network traffic anomaly detection based on KL distance[J]. Computer Engineering, 2019, 45(4): 108-113, 118.)

[10] JIN W, TUNG A K H, HAN J, et al. Ranking outliers using symmetric neighborhood relationship [C]// Proceedings of the 2006 Pacific-Asia Conference on Knowledge Discovery and Data Mining, LNCS 3918. Berlin: Springer, 2006: 577-593.

[11] 阮嘉琨,蔡延光,乐冰. 基于DBSCAN密度聚类算法的高速公路交通流异常数据检测[J]. 工业控制计算机, 2019, 32(7): 92-94. (RUAN J K, CAI Y G, LE B. Highway traffic anomaly data detection based on DBSCAN density clustering algorithm[J]. Industrial Control Computer, 2019, 32(7):92-94.)

[12] PAN X, TAN J, KAVULYA S, et al. GaneSHA: black-Box diagnosis of MapReduce systems[J]. ACM SIGMETRICS Performance Evaluation Review, 2010, 37(3):8-13.

[13] GUPTA M, SHARMA A B, CHEN H, et al. Context-aware time series anomaly detection for complex systems[EB/OL]. [2019-05-20]. https://www. microsoft. com/en-us/research/wp-content/uploads/2013/01/gupta13_sdma. pdf.

[14] CHANDOLA V, MITHAL V, KUMAR V. Comparative evaluation of anomaly detection techniques for sequence data[C]// Proceedings of the 8th IEEE International Conference on Data Mining. Piscataway: IEEE, 2008:743-748.

[15] SUN P, CHAWLA S, ARUNASALAM. Mining for outliers in sequential databases[EB/OL]. [2019-05-20]. https://archive. siam. org/meetings/sdm06/proceedings/009sunp. pdf.

[16] CHEN P Y, YANG S, MCCANN J A. Distributed real-time anomaly detection in networked industrial sensing systems[J]. IEEE Transactions on Industrial Electronics, 2015, 62(6):3832-3842.

[17] FARIA E R, GAMA J, CARVALHO A C P L F. Novelty detection algorithm for data streams multi-class problems[C]// Proceedings of the 28th Annual ACM Symposium on Applied Computing. New York: ACM, 2013:795-800.

[18] YU Y, GUO S, LAN S, et al. Anomaly intrusion detection for evolving data stream based on semi-supervised learning[C]// Proceedings of the 15th International Conference on Neural Information Processing, LNCS 5506. Berlin: Springer, 2008:571-578.

[19] AHMAD S, LAVIN A, PURDY S, et al. Unsupervised real-time anomaly detection for streaming data[J]. Neurocomputing, 2017, 262: 134-147.

[20] SZMIT M, SZMIT A. Use of holt-winters method in the analysis of network traffic: case study[C]// Proceedings of the 2011 International Conference on Computer Networks, CCIS 160. Berlin: Springer, 2011:224-231.

[21] BASSEVILLE M, NIKIFOROV I V. Detection of Abrupt Change Theory and Application[M]. Upper Saddle River: Prentice Hall, 1993: 23-26.

[22] BIANCO A M, BEN M G, MARTÍNEZ E J, et al. Outlier detection in regression models with ARIMA errors using robust estimates [J]. Journal of Forecasting, 2001, 20(8):565-579.

[23] GREFF K, SRIVASTAVA R K, KOUTNÍK J, et al. LSTM: a search space odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(10):2222-2232.

[24] UCI Machine Learning Repository. Condition monitoring of hydraulic systems[DB/OL]. [2019-07-20] http://archive. ics. uci. edu/ml/datasets/Condition+monitoring+of+hydraulic+systems.

[25] ZHANG Y, SZABO C, SHENG Q Z. Cleaning environmental sensing data streams based on individual sensor reliability [C]// Proceedings of the 2014 International Conference on Web Information Systems Engineering, LNCS 8787. Cham: Springer, 2014: 405- 414.

[26] ZIMEK A, CAMPELLO R J G B, SANDER J. Ensembles for unsupervised outlier detection: challenges and research questions a position paper[J]. ACM SIGKDD Explorations Newsletter, 2014, 15(1):11-22.

QIU Yuan, born in 1995, M. S. candidate. Her research interests include anomaly detection, deep learning.

CHANG Xiangmao, born in 1982, Ph. D., associate professor. His research interests include internet of things, intelligent health monitoring based on wearable devices, sensory data processing and analysis of machine learning algorithms.

QIU Qian, born in 1995, M. S. candidate. Her research interests include social network, deep learning.

PENG Cheng, born in 1995, M. S. candidate. His research interests include state detection, deep learning.

SU Shanting, born in 1994, M. S. candidate. Her research interests include fault detection, machine learning.