



文章编号:1001-9081(2019)03-0924-06

DOI:10.11772/j.issn.1001-9081.2018081681

基于双向 LSTM 的 Seq2Seq 模型在加油站时序数据异常检测中的应用

陶 涛^{1,2,3}, 周 喜^{1,3*}, 马 博^{1,3}, 赵 凡^{1,3}

(1. 中国科学院 新疆理化技术研究所, 乌鲁木齐 830011; 2. 中国科学院大学, 北京 100049;

3. 新疆理化技术研究所 新疆民族语音语言信息处理实验室, 乌鲁木齐 830011)

(* 通信作者电子邮箱 zhousx@ms.xjb.ac.cn)

摘要: 加油时序数据包含加油行为的多维信息,但是指定加油站点数据较为稀疏,现有成熟的数据异常检测算法存在挖掘较多假性异常点以及遗漏较多真实异常点的缺陷,并不适用于挖掘加油站时序数据。提出一种基于深度学习的异常检测方法识别加油异常车辆,首先通过自动编码器对加油站点采集到的相关数据进行特征提取,然后采用嵌入双向长短期记忆(Bi-LSTM)的 Seq2Seq 模型对加油行为进行预测,最后通过比较预测值和原始值来定义异常点的阈值。通过在加油数据集以及信用卡欺诈数据集上的实验验证了该方法的有效性,并且相对于现有方法在加油数据集上均方根误差(RMSE)降低了 21.1%,在信用卡欺诈数据集上检测异常的准确率提高了 1.4%。因此,提出的模型可以有效应用于加油行为异常的车辆检测,从而提高加油站的管理和运营效率。

关键词: 加油站时序数据; 深度学习; Seq2Seq; 双向长短期记忆; 异常检测

中图分类号: TP391.4 **文献标志码:** A

Abnormal time series data detection of gas station by Seq2Seq model based on bidirectional long short-term memory

TAO Tao^{1,2,3}, ZHOU Xi^{1,3*}, MA Bo^{1,3}, ZHAO Fan^{1,3}

(1. Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi Xinjiang 830011, China;

2. University of Chinese Academy of Sciences, Beijing 100049, China;

3. Xinjiang Laboratory of Minority Speech and Language Information Processing, Xinjiang Technical Institute of Physics and Chemistry, Xinjiang Urumqi 830011, China)

Abstract: Time series data of gas station contains multi-dimensional information of fueling behavior, but the data of specific gas station are sparse. The existing abnormal data detection algorithms are not suitable for gas station time series data, because many pseudo outliers are mined and many real abnormal points are missed. To solve the problems, an abnormal detection method based on deep learning was proposed to detect vehicles with abnormal fueling. Firstly, feature extraction was performed on data collected from the gas station through an automatic encoder. Then, a deep learning model Seq2Seq with embedding Bidirectional Long Short-Term Memory (Bi-LSTM) was used to predict the fueling behavior. Finally, the threshold of outliers was defined by comparing the predicted value and the original value. The experiments on a fueling dataset and a credit card fraud dataset verify the effectiveness of the proposed method. Compared with the existing methods, the Root Mean Squared Error (RMSE) of the proposed method is decreased by 21.1% on the fueling dataset, and abnormal detection accuracy of the proposed method is improved by 1.4% on the credit card fraud dataset. Therefore, the proposed method can be applied to detect vehicles with abnormal fueling behavior, improving the management and operational efficiency of gas station.

Key words: gas station time-series data; deep learning; Seq2Seq; Bidirectional Long Short-Term Memory (Bi-LSTM); outlier detection

0 引言

加油数据采集系统的广泛使用产生了大量的加油数据,包含了丰富有价值的信息。然而数据采集系统发生故障或者人为记录失误会使得加油站点数据采集完成后产生异常数据,异常数据包括加油量异常以及加油行为异常等,如何有效

地检测异常数据点对于提高加油站的管理和运营效率有着重要意义。目前关于数据异常检测的方法有很多种,并且已经取得一定的成果。然而对于指定加油站点加油数据排列较为稀疏且没有异常数据标签,现有成熟的异常检测算法在加油站时序数据上表现不尽如人意,甚至有的算法并不适用,在进行异常数据检测时存在挖掘较多假性异常点以及遗漏较多真

收稿日期:2018-08-14;修回日期:2018-09-13;录用日期:2018-09-18。

基金项目:新疆维吾尔自治区高层次人才引进工程资助项目(Y639401201);中国科学院西部之光项目(2016-QNXZ-A-3)。

作者简介:陶涛(1994—),男,贵州毕节人,硕士研究生,主要研究方向:大数据分析、数据挖掘; 周喜(1978—),男,湖南双峰人,研究员,博士,CCF 会员,主要研究方向:物联网、大数据分析; 马博(1984—),男,辽宁鞍山人,副研究员,博士,CCF 会员,主要研究方向:数据分析与知识发现、机器学习; 赵凡(1980—),男,山西介休人,副研究员,博士研究生,CCF 会员,主要研究方向:信息安全、大数据分析。



实异常点的缺陷。为了更为有效、准确地挖掘加油站时序数据中的异常值,本文针对加油站时序数据提出一种基于深度学习的方法——TS-DL(Time-Series based Deep Learning)来检测异常。首先,利用深度学习的特征学习以及信息记忆能力来对输入时间序列数据进行预测,然后再比较原始数据和预测数据的差异值,依据 3σ 准则^[1]设定阈值,最终检测到异常点。本文的主要工作分为三个部分:1)利用自动编码机(AutoEncoder)从原始数据集中提取有效特征。2)利用嵌入双向长期记忆(Bidirectional Long Short-Term Memory, Bi-LSTM)^[2]的Seq2Seq^[3]模型预测包含1)中特征以及附加特征的时间序列数据。3)通过比较原始数据和2)中预测数据的差异值来挖掘异常点。图1为本文方法的总体流程。

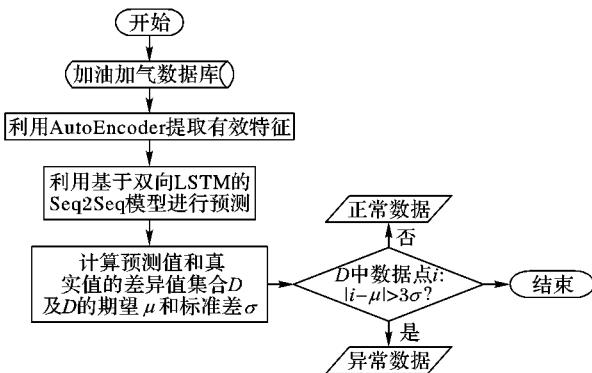


图1 本文方法的总体流程

Fig. 1 Overall procedure of the proposed method

1 相关工作

目前关于异常检测的算法主要可以分成四类,分别为:基于统计的模型、基于距离的模型、基于线性变换的模型以及基于非线性变换的模型。

1.1 基于统计的模型

基于统计的模型基于以下关键性假设:正常数据实例出现在随机模型的高概率分布区域,而异常数据出现在随机模型的低概率分布区域^[4]。该方法使用一些分布模型来拟合数据,并且认为分布在边缘处的数据是异常的。对于一维数据集 $D = \{x_1, x_2, \dots, x_n\}$,如果数据集可以使用高斯分布进行拟合,则根据该数据,可以计算高斯分布的两个参数期望 $\mu = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ 以及方差 $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ 。假如一个实例 $x \notin (\mu - 3\sigma, \mu + 3\sigma)$ 则认为它为异常点。对于多维数据集 $D = \{x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}) : 1 \leq i \leq m\}$,首先通过 χ^2 检验计算数据集每一维的期望 $E_k = \sum_{i=1}^m x_k p_k$,然后再计算数据实例的 $\chi^2 = \sum_{k=1}^n \frac{(x_k - E_k)^2}{E_k}$,如果 χ^2 比较大,则表明该数据实例为异常。由于加油站时序数据维度较高,数据分布较为稀疏,此时对数据的分布难以确保成立,故基于统计的模型不适用于加油站时序数据。

1.2 基于距离的模型

基于距离的模型^[5]主要分为基于角度、基于距离以及基于密度的方法。基于角度的方法认为相对聚集的数据实例彼此间的角度要远远小于相对分散的数据实例,而分散的数据

实例可以认为是异常的,所以可以利用数据实例间的角度进行异常检测。基于距离的方法相比基于角度的方法更直观,该类方法利用数据实例的距离衡量其聚集程度,离群点即为异常点。此类方法衍生出许多方法,包括基于聚类、基于分类以及基于划分的方法等。基于密度的方法如 LOF(Loacl Outlier Factor)算法比较数据实例及其周围 k 个最邻近数据实例的密度,如果两者差异越大,则意味着该数据实例为异常数据的可能性越大。在加油站时序数据中正常数据没有足够的邻居或者异常点有很多邻居,另外对于时序数据定义数据之间的距离也会很困难,故基于距离的模型不适用于加油站时序数据。

1.3 基于线性变换的模型

基于线性变换的模型基于以下关键假设:数据可嵌入到低维子空间中,其中正常情况和异常情况存在显著不同^[6]。检测异常值的线性模型可分为两类:第一类模型主要使用统计回归建模^[7];第二类模型使用主成分分析(Principal Component Analysis, PCA)来确定投影的低维子空间^[8]。在加油站时序数据中正常点和异常点的界限并不是很明确,正常数据点的表现不断变化,异常点难以确认,故基于线性变换的模型不适用于加油站时序数据。

1.4 基于非线性变换的模型

基于非线性的模型主要针对多维大型数据库的异常检测问题,通过神经网络较强的学习能力来计算异常数据的偏离度。在过去的几年里,深度学习在数据挖掘和分析方面取得了很大的进展,并提出了各种有效模型。对于序列数据往往采用基于循环神经网络(Recurrent Neural Network, RNN)的方法^[9],这类方法首先采用正常数据对网络进行训练,对于待检测数据实例,利用重构误差(reconstruction error)作为异常数据的度量值:

$$\text{reconstruction_error}(x^{(i)}) = \frac{1}{n} \sum_{j=1}^n (x_j^{(i)} - o_j^{(i)})^2$$

综上所述,神经网络具备的强特征提取和信息记忆能力更适用于加油站时序数据,利用神经网络构建契合的模型能够较好地完成加油站时序数据异常检测工作。

2 异常检测模型

2.1 问题定义及分析

加油站数据集可以用加油对象集合: $O_{DB} = \{o_1, o_2, \dots, o_n\}$ 来表示,其中一条加油数据 o_i 包括了多个特征信息,例如加油时间、一次加油油量、汽油类型、加油车辆类型等,这些数据经过各个站点采集后采用非结构化的方式存储在数据库中。对于指定一个区域内的加油站点,根据可视化分析发现其加油行为具有一定的周期性和趋势性。

RNN具有有限的短期记忆优势,所以RNN作为训练时序数据的首选神经网络。然而RNN只能学习到一定间隔时间序列信息,当序列数据超过一定长度时,利用RNN训练数据会出现严重的梯度消失问题而导致训练停止^[10]。由于 O_{DB} 中时间序列较长,在这种情况下RNN无法有效利用这些长序列历史信息,即无法学习到长依赖的特征。为了尽可能有效学习到 O_{DB} 中时间序列信息,本文提出一种嵌入Bi-LSTM的Seq2Seq模型,其中Seq2Seq是一种Encoder-Decoder结构



的网络模型,其输入序列和输出序列都是可变长度的,基于这样的机制,嵌入性能较好的 Bi-LSTM 进行预测,相比于单纯使用长短期记忆(Long Short-Term Memory, LSTM)进行预测,此方法具有较优的效率。

结合上述分析,本文为了有效挖掘 O_{DB} 中异常模式提出一种基于深度学习的异常检测方法,该方法首先通过自动编码器(AutoEncoder)对加油站点采集到的相关数据进行特征提取,然后采用嵌入 Bi-LSTM 的 Seq2Seq 模型对加油行为进行预测,最后依照 3σ 准则比较预测值和原始值来定义异常点的阈值。

2.2 特征提取

由于原始数据中数据维度大且较为稀疏,首先采用自动编码机^[11]对其进行特征提取。自动编码机是神经网络的一种,经过训练后能尝试将输入复制到输出。但是为了使之能够学习到有效的特征,通过强加一些约束,使自动编码机只能近似地复制,从而能够学习到数据的有用特性达到数据降维的效果。

如图 2,通过限制 L_2 层(隐藏层)的维度,使其比 L_1 层(输入层)维度低,这样强制自动编码机捕捉训练数据中最显著的特征,从而达到特征提取的目的。

算法伪代码:数据特征提取。

自动编码机首先通过预训练得到的权重矩阵 \mathbf{W} 对输入进行压缩编码,经激活函数后再解码恢复数据以期望输出等于输入,通过迭代的训练待整个模型收敛时得到训练完成的自动编码机。

输入:原始数据集 x_i 以及数据属性值 y_i (为 x_i 的权值)

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$$

输出:训练完成的自动编码机(参数 \mathbf{W} 和 \mathbf{b})

定义 a_j^l 为第 L 层第 j 个单元节点激活量; s_l 为第 L 层节点数量; f 为激活函数 sigmoid; \mathbf{W} 为权重矩阵; \mathbf{b} 为偏置向量; z_j^l 为第 L 层节点 j 激活量的输入; 损失函数

$$J(\mathbf{W}, \mathbf{b}) = \left[\frac{1}{m} \sum_{i=1}^m \frac{1}{2} h_{\mathbf{W}, \mathbf{b}}(x^{(i)} - y^{(i)})^2 \right] + \frac{\lambda}{2} \sum_{l=1}^{n_L-1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_{l+1}} (\mathbf{W}_{ji}^l)^2$$

其中: $h_{\mathbf{W}, \mathbf{b}}(x) = f(z_j^l) = f\left(\sum_i^{s_l-1} (\mathbf{W}_{ji}^{l-1} * a_i^{l-1}) + b_j^{l-1}\right)$, 损失函数第一项为平均平方和误差,第二项为正则项,正则项的添加是为了减少权值的量级以防止训练过度拟合。采用梯度下降法训练使得 $J(\mathbf{W}, \mathbf{b})$ 最小:

1) $\Delta \mathbf{W}^{(l)} := 0, \Delta \mathbf{b}^{(l)} := 0$ // 对于每一层参数进行初始化

2) for each epoch:

令: $\Delta \mathbf{W}^{(l)} = \Delta \mathbf{W}^{(l)} + \nabla_{\mathbf{W}^{(l)}} J(\mathbf{W}, \mathbf{b}; x, y)$

$\Delta \mathbf{b}^{(l)} = \Delta \mathbf{b}^{(l)} + \nabla_{\mathbf{b}^{(l)}} J(\mathbf{W}, \mathbf{b}; x, y)$

采用反向传播计算:

$$\nabla_{\mathbf{W}^{(l)}} J(\mathbf{W}, \mathbf{b}; x, y) = \left[\frac{1}{m} \sum_{i=1}^m \left(\frac{\partial}{\partial \mathbf{W}_{ij}^l} J(\mathbf{W}, \mathbf{b}; x^{(i)}, y^{(i)}) \right) \right]$$

$$\nabla_{\mathbf{b}^{(l)}} J(\mathbf{W}, \mathbf{b}; x, y) = \left[\frac{1}{m} \sum_{i=1}^m \left(\frac{\partial}{\partial \mathbf{b}_i^l} J(\mathbf{W}, \mathbf{b}; x^{(i)}, y^{(i)}) \right) \right]$$

更新参数: $\mathbf{W}^{(l)} = \mathbf{W}^{(l)} - \alpha \left[\frac{1}{m} \Delta \mathbf{W}^{(l)} + \lambda \mathbf{W}^{(l)} \right]$

$$\mathbf{b}^{(l)} = \mathbf{b}^{(l)} - \alpha \left[\frac{1}{m} \Delta \mathbf{b}^{(l)} \right] \quad // \alpha \text{ 为学习率}$$

3) 迭代进行 2) 中的参数计算,直至其参数 \mathbf{W} 和 \mathbf{b} 迭代更新到不再发生变化或者变化量极小即收敛时停止训练。

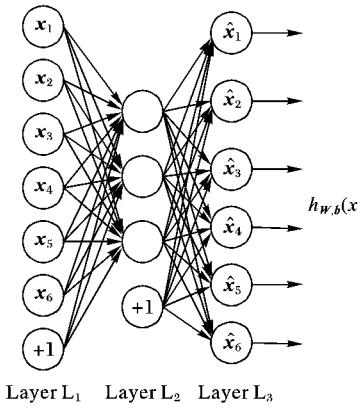


图 2 自动编码机结构

Fig. 2 Structure of AutoEncoder

2.3 加油对象预测

为了更为准确地完成本文核心的预测任务,主要通过对 Seq2Seq 模型^[12]进行变换来实现。Seq2Seq 模型的核心思想是把一个语言序列翻译成另外一种语言序列,整个处理过程是通过使用 RNN 将一个序列作为输入映射为另外一个输出序列。而 RNN 在处理时序数据上往往会过于依赖邻近点数据而忽略长距离的信息,后来出现的 LSTM 在 RNN 上作了改进,使其能够捕捉到更长距离的信息,从而学习到长依赖的特征^[13]。但是不论是 LSTM 还是 RNN,在进行预测时都是从前向后进行的,因此后面的数据点会比前面的更加重要,这样往往回遗漏许多长关联数据点的信息。而 Bi-LSTM^[14] 的出现改善了这种缺陷,其基本机制是对于一个训练序列进行向前和向后两次 LSTM 训练,而且它们都连接着一个输出层,从而提供给输出层输入序列中每一个点完整的过去和未来的上下文信息,从而构建了基于 Bi-LSTM 的 Seq2Seq 预测模型(记为 BL-Seq2Seq 模型)。

此外对于长时间序列数据,为了加强模型的记忆能力,本文没有采用 Bahdanau 或 Luong 注意力机制^[15],因为经典的注意力机制在每个预测步长上使用所有的历史数据点从头开始计算,这样对于长时间序列数据来说计算复杂度是无法承受的。因此取而代之的方案是将时间序列中重要的数据点(节假日、双休日等)作为编码器和解码器的附加特征和处理后的数据一起放入模型中进行训练,在采样的数据序列上进行实验,文献[16]结果表明这样的方法能够有效加强模型的记忆能力。

如图 3,在本文的预测模型 BL-Seq2Seq 中,包含附加特征的序列数据(x_1, x_2, \dots, x_n)进入编码器 Encoder 中,完成编码得到语义向量 e 。然后将 e 放入解码器 Decoder 中,解码器根据上一个时刻的输出会作为当前时刻的输入,依此循环完成预测。

算法伪代码:加油数据预测。

在预测模型 BL-Seq2Seq 中,输入经过预处理过的时间序列数据,首先进入具备 Bi-LSTM 结构的编码层中,通过编码完



成得到语义向量,然后语义向量进入相似结构的解码层中解码完成,最终根据局部最优算法计算得到预测概率最大的点,并且依次循环预测得到输出的预测数据序列。

输入:包含附加特征的序列 $S_1 = (x_1, x_2, \dots, x_n)$

输出:预测数据序列 $S_1' = (y_1, y_2, \dots, y_n)$

1) 单向 LSTM。

定义 W_{ix} 为各个权重矩阵; b 为偏置向量; σ 为 sigmoid 函数; c 为 cell 状态更新向量; m 为输出向量; \odot 为点乘; g, h 为 cell 的输入输出激活函数,一般为 tanh; φ 为最终输出激活函数,一般为 softmax。

$$i_t = \sigma(W_{ix} * x_t + W_{im} * m_{t-1} + W_{ic} * c_{t-1} + b_i) \quad // \text{输入门}$$

$$f_t = \sigma(W_{fx} * x_t + W_{fm} * m_{t-1} + W_{fc} * c_{t-1} + b_f) \quad // \text{遗忘门}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g(W_{cx} * x_t + W_{cm} * m_{t-1} + b_c) \quad // \text{cell 状态更新}$$

$$o_t = \sigma(W_{ox} * x_t + W_{om} * m_{t-1} + W_{oc} * c_t + b_o) \quad // \text{输出门}$$

$$m_t = o_t \odot h(c_t) \quad // \text{输出向量}$$

$$y_t = \varphi(W_{fx} * m_t + b_y) \quad // \text{最终输出}$$

2) Bi-LSTM。

定义 U, V, W 为各个权重矩阵; f, g 为 LSTM 激活函数

$$s_t = f(U * x_t + W * s_{t-1}) \quad // \text{正向计算 LSTM 隐藏层状态}$$

$$s'_{t'} = f(U' * x_t + W' * s_{t+1}) \quad // \text{反向计算 LSTM 隐藏层状态}$$

$$o_t = g(V * s_t + V' * s'_{t'}) \quad // \text{最终输出取决于 } s \text{ 和 } s'$$

3) BL-Seq2Seq。

Encoder: $\quad // \text{编码过程}$

$$h_t = f(x_t, h_{t-1}) \quad // \text{当前节点状态}$$

$$e = \varphi(h_1, h_2, \dots, h_n) \quad // \text{生成语义向量}$$

Decoder: $\quad // \text{解码过程}$

$$h_t = f(h_{t-1}, y_{t-1}, e) \quad // \text{计算当前节点状态}$$

$$P(y_t | y_{t-1}, \dots, y_1, e) = g(h_t, y_{t-1}, e) \quad // \text{采用局部最优解算法计算对于预测概率最大的点}$$

Final:

$$y_1 = f(e), y_2 = f(e, y_1), y_3 = f(e, y_1, y_2), \dots, y_n = f(e,$$

$$y_1, y_2, \dots, y_{n-1})$$

End

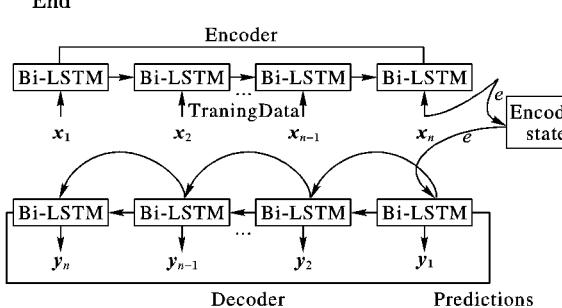


图3 预测模型核心图

Fig. 3 Core diagram of prediction model

2.4 异常对象挖掘

通过2.3节可以得到预测的数据集合 S' ,显然易得预测数据和实际数据的差异值集合 $D = |S' - S|$,再计算 D 的均值 μ 和标准差 σ ,并将 D 中数据拟合到正态分布上,最终定义 D 中数据 i 满足 $|i - \mu| > 3\sigma$ 条件的数据点为异常点。

算法伪代码:异常对象挖掘。

输入:原始数据序列 S ,预测数据序列 S' ;

输出:异常点集合 A 。

```
D = |S' - S| // 预测值和实际值的差值
For i in D:
    μ = x̄ = 1/n ∑i=1n xi // 计算 D 的均值
    σ = √( ∑i=1n (xi - x̄)2 ) / n // 计算 D 的标准差
    x̂ = len(D)
    ̂y = norm(x̂, μ, σ) // 拟合到正态分布上
    If |i - μ| > 3σ // 判断是否是异常点
        add i to A
return A
```

3 实验与分析

3.1 实验配置

为了验证该方法对异常对象挖掘的准确性和有效性,本文在两个数据集上进行了实验,两个数据集分别为中国某省份汽车加油数据集以及信用卡欺诈检测数据集^[17]。前者为无异常标签标注的时序数据集,后者为带异常标签标注的公开数据集。实验机器系统为Win7 64位,CPU型号为Intel Core i7-4720HQ CPU @ 2.60 GHz,内存8 GB,python版本为3.6,keras版本为2.0.8,使用的数据库为MongoDB3.0。

3.2 加油数据集

采用的是数据集是中国某省的各个加油站点的加油数据,通过进行融合、清洗,然后再对数值型特征进行归一化,非数值型特征进行数字编码后再归一化处理,此外将时间序列中属于节假日和双休日的数据作为附加特征标注后放入训练数据中,这样最终得到可靠的实验数据。为验证算法有效性,将预处理后的数据分别取80%作为训练集,将剩下20%作为测试集。

为了验证本文核心预测模型(BL-Seq2Seq)在预处理后的数据集上的性能,通过与标准LSTM模型进行比较评估。采用的评价函数主要有如下两个。

1) 模型训练过程中的损失函数。本文采用神经网络模型训练中常用的均方误差(Mean Squared Error, MSE),其具体公式如下:

$$MSE = \frac{1}{m} \sum_{i=1}^m (h(x_i) - y_{(i)})^2$$

2) 加载训练完成的模型进行预测时采用均方根误差(Root Mean Square Error, RMSE)。其具体公式如下:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (h(x_i) - y_{(i)})^2}$$

其中 $h(x_i)$ 和 y_i 分别表示预测值和实际值, m 代表样本总数。

3.3 信用卡欺诈数据集

信用卡欺诈数据集是公开数据集,包含两天各个时间点



的284 807笔交易记录,有492笔交易已标注为欺诈行为。影响欺诈因素包含有28个数值型变量 v_1, v_2, \dots, v_{28} ,另外两列数据是交易金额Amount和欺诈标签Class。经验证原数据集数据项完整且影响欺诈因素的28个数值型变量都已通过PCA变换处理完毕,故只需对交易金额进行归一化处理。为验证算法有效性,将处理后的数据分别从正常值取80%作为训练集,将剩下20%的正常值和所有异常值(欺诈行为)作为测试集。

为了验证本文算法(TS-DL)的性能,通过与经验证在此数据集上效果好的逻辑回归(LogisticRegression)模型^[18]以及雅虎(Yahoo)大规模时序数据自动异常检测架构(Extensible Generic Anomaly Detection System, EGADS)^[19]中的主要预测模型进行比较评估。实验常用的评价标准用准确率和召回率以及F1score等;但是在本文异常检测的场景中,实验所用数据集为非均衡数据集,且正负样本数量差距较大,即异常对象和正常对象比例差异非常大,故此时上述评价标准无法全面地展示算法性能。Shi等^[20]指出马修斯系数(Matthews Correlation Coefficient, MCC)能够有效衡量不平衡数据集,所以本文将MCC作为主要性能衡量指标。具体公式如下:

$$MCC =$$

$$\frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

其中: TP (True Positive)表示异常对象挖掘表示为异常的样本数; TN (True Negative)表示正常对象挖掘表示为正常的样本数; FP (False Positive)表示正常对象挖掘表示为异常的样本数; FN (False Negative)表示异常对象挖掘表示为正常的样本数。

此外为了衡量本文算法的泛化性能,采用ROC(Receiver Operating Characteristic)曲线^[21]来评估。ROC曲线的纵轴是真正例率(True Positive Rate, TPR),横轴是假正例率(False Positive Rate, FPR),二者分别定义为:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

3.4 实验结果与分析

3.4.1 加油数据集实验

1) 模型训练过程中的损失函数对比。

实验在采样数据点上分别比较LSTM向前传播(LSTM_forw)、LSTM向后传播(LSTM_back)以及本文模型BL-Seq2Seq在迭代训练过程中损失函数loss的平均值如表1所示。

表1 LSTM和BL-Seq2Seq损失函数对比

Tab. 1 Loss function comparison between LSTM and BL-Seq2Seq

模型	loss 平均值	模型	loss 平均值
LSTM_forw	0.3947	BL-Seq2Seq	0.2450
LSTM_back	0.4721		

表1中loss平均值表示模型在经过250轮训练后得到的损失函数平均值。由表1易知,相对于LSTM的前后向传播,本文模型BL-Seq2Seq在训练过程中具有较低的损失函数,说

明BL-Seq2Seq具有较好的模型拟合效果。

2) 进行预测时的均方根误差对比。

在采样的数据点上分别对EGADS中移动平均模型MovingAverageModel、统计模型NaiveForecastingModel、回归模型RegressionModel、季节模型OlympicModel、指数平滑模型DoubleExponentialSmoothingModel、LSTM模型以及BL-Seq2Seq模型进行预测实验,用RMSE去度量其预测效果,结果如表2所示。

表2 不同模型RMSE对比

Tab. 2 RMSE comparison of different models

模型	RMSE
MovingAverageModel	32.23
NaiveForecastingModel	36.21
RegressionModel	33.59
OlympicModel	32.43
DoubleExponentialSmoothingModel	23.50
LSTM	18.16
BL-Seq2Seq	14.33

实验结果表明:相比于EGADS中经典的预测模型,LSTM模型和本文模型BL-Seq2Seq明显具有较低的RMSE,说明采用神经网络的两种模型大大降低了预测误差。而本文模型BL-Seq2Seq相比于目前成熟的LSTM模型,预测误差RMSE降低了21.1%,说明本文模型BL-Seq2Seq在对采样的加油数据进行预测时相对于当前性能较好的LSTM模型具有较低的预测误差,证明本文模型能够有效提升一定的预测准确度。

3.4.2 信用卡欺诈数据集实验

1) ROC曲线图对比。

在本实验中欺诈行为挖掘的准确性由TP和TN占总样本的比例决定,此比例越高代表准确率越高。为比较两个算法的准确性,分别绘制二者ROC曲线如图4所示。

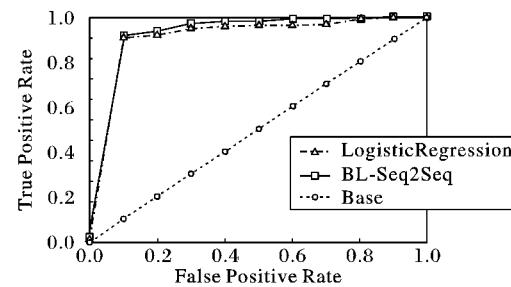


图4 LogisticRegression和BL-Seq2Seq ROC曲线对比

Fig. 4 Comparison of ROC curve between

LogisticRegression and BL-Seq2Seq

由ROC曲线的性质:ROC曲线越靠近左上角所代表的分类结果越准确,此外亦可通过分别计算各个模型的ROC曲线下的面积(Area Under the Curve, AUC)进行比较,AUC值越大表示模型的分类性能越好。由图4可知,在相同的数据集上,本文提出的模型BL-Seq2Seq相对于逻辑回归(LogisticRegression)模型ROC曲线更靠近左上角,且经计算可得逻辑回归模型的AUC值为0.9458而BL-Seq2Seq模型的AUC值为0.9602,说明本文提出的模型性能更好。

2) MCC对比。

MCC针对不平衡的数据集具有较好的评估效果,实验结



果可知, LogisticRegression 和 BL-Seq2Seq 的马修斯系数(MCC) 分别为 0.230 0 和 0.359 7。BL-Seq2Seq 相比于 LogisticRegression 具有较高的 MCC, 证明本文算法能够有效地检测到欺诈行为, 并且提升了一定的检测准确度。

4 结语

本文鉴于循环神经网络在长时间序列预测时存在的缺陷, 提出一种采用嵌入 Bi-LSTM 的 Seq2Seq 模型并将重要数据点作为 Seq2Seq 的附加特征进行预测从而检测数据异常的方法。该方法首先对数据集中高维特征通过自动编码机进行特征提取, 然后将处理后的数据及附加特征一起放入嵌入 Bi-LSTM 的 Seq2Seq 模型进行训练, 接着加载训练好的模型进行相应的数据预测, 最后比较预测值与真实值的差异值并将其拟合到正态分布上通过 3σ 准则检测异常。在加油数据集以及信用卡欺诈数据集上的实验说明了本文方法有效且对于现有较好算法有了一定程度的改进。但是在实验过程中将差异值直接拟合到数据分布上的方法较为简单, 后续的研究中将探索更多的异常检测方法。另外数据集中总体数据量巨大, 在后续工作中将采用服务器多 GPU 进行并行化处理。

参考文献 (References)

- [1] ROUSSEEUW P J, LEROY A M. Robust Regression and Outlier Detection [M]. New York: John Wiley & Sons, 2005: 254 – 255.
- [2] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. [2015-08-09]. <https://arxiv.org/pdf/1508.01991.pdf>.
- [3] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks [C]// NIPS 2014: Proceedings of the 2014 Advances in Neural Information Processing Systems 27. Montréal: [s. n.], 2014: 3104 – 3112.
- [4] 严宏, 杨波, 杨红雨. 基于异方差高斯过程的时间序列数据离群点检测[J]. 计算机应用, 2018, 38(5): 1346 – 1352. (YAN H, YANG B, YANG H Y. Outlier detection in time series data based on heteroscedastic Gaussian processes [J]. Journal of Computer Applications, 2018, 38(5): 1346 – 1352.)
- [5] 陈斌, 陈松灿, 潘志松, 等. 异常检测综述[J]. 山东大学学报(工学版), 2009, 39(6): 13 – 23. (CHEN B, CHEN S C, PAN Z S, et al. Survey of outlier detection technologies [J]. Journal of Shandong University (Engineering Science), 2009, 39(6): 13 – 23.)
- [6] HUANG T, ZHU Y, WU Y, et al. Anomaly detection and identification scheme for VM live migration in cloud infrastructure [J]. Future Generation Computer Systems, 2016, 56(C): 736 – 745.
- [7] WANG T, LI Z. Outlier detection in high-dimensional regression model [J]. Communications in Statistics, 2016, 46(14): 6947 – 6958.
- [8] 鲍苏宁, 张磊, 杨光. 基于核主成分分析的异常轨迹检测方法 [J]. 计算机应用, 2014, 34(7): 2107 – 2110. (BAO S N, ZHANG L, YANG G. Trajectory outlier detection method based on kernel principal component analysis [J]. Journal of Computer Applications, 2014, 34(7): 2107 – 2110.)
- [9] SHIPMON D T, GUREVITCH J M, PISELLI P M, et al. Time series anomaly detection: detection of anomalous drops with limited features and sparse examples in noisy highly periodic data [EB/OL]. [2017-08-11]. <http://cn.arxiv.org/ftp/arxiv/papers/1708/1708.03665.pdf>.
- [10] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2016-05-19]. <https://arxiv.org/pdf/1409.0473.pdf>.
- [11] FARIA G, DORMIDO-CANTO S, VEGA J, et al. Automatic feature extraction in large fusion databases by using deep learning approach [J]. Fusion Engineering and Design, 2016, 112: 979 – 983.
- [12] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks [EB/OL]. [2018-07-10]. <https://arxiv.org/pdf/1409.3215.pdf>
- [13] ZHENG J, XU C, ZHANG Z, et al. Electric load forecasting in smart grids using long-short-term-memory based recurrent neural network [C]// CISS 2017: Proceedings of the 2017 51st Annual Conference on Information Sciences and Systems. Piscataway, NJ: IEEE, 2017: 1 – 6.
- [14] TAI K S, SOCHER R, MANNING C D. Improved semantic representations from tree-structured long short-term memory networks [EB/OL]. [2018-05-30]. <https://arxiv.org/pdf/1503.00075.pdf>.
- [15] CHO K, van MERRIENBOER B, GULCEHRE C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation [EB/OL]. [2017-09-03]. <https://arxiv.org/pdf/1406.1078.pdf>.
- [16] MADLENÁK R, MADLENÁKOVÁ L, SVADLENKA L, et al. Analysis of website traffic dependence on use of selected Internet marketing tools [J]. Procedia Economics and Finance, 2015, 23: 123 – 128.
- [17] AGNIHOTRI M. Credit card fraud detection [DB/OL]. [2017-04-27]. <https://www.ushuji.com/financial/296.html>.
- [18] TSANGARATOS P, ILIA I. Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: the influence of models complexity and training dataset size [J]. Catena, 2016, 145: 164 – 179.
- [19] LAPTEV N, AMIZADEH S, FLINT I. Generic and scalable framework for automated time-series anomaly detection [C]// KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2015: 1939 – 1947.
- [20] SHI Y, EBERHART R C. Empirical study of particle swarm optimization [C]// CEC '99: Proceedings of the 1999 Congress on Evolutionary Computation. Piscataway, NJ: IEEE, 1999, 3: 1945 – 1950.
- [21] 周志华. 机器学习: = Machine learning[M]. 北京: 清华大学出版社, 2016: 33 – 36. (ZHOU Z H. Machine learning: = Machine learning [M]. Beijing: Tsinghua University Press, 2016: 33 – 36.)

This work is partially supported by the Program of Introducing High-Level Talents of Xinjiang(Y639401201), the West Light Foundation of Chinese Academy of Sciences (2016-QNXZ-A-3).

TAO Tao, born in 1994, M. S. candidate. His research interests include big data analysis, data mining.

ZHOU Xi, born in 1978, Ph. D., research fellow. His research interests include Internet of things, big data analysis.

MA Bo, born in 1984, Ph. D., associate research fellow. His research interests include data analysis and knowledge discovery, machine learning.

ZHAO Fan, born in 1980, Ph. D. candidate, associate research fellow. His research interests include information security, big data analysis.