基于自注意力和 Bi-LSTM 的业务流程异常检测模型

陈国威[1],卢 可[1,2]
1. 安徽理工大学 数学与大数据学院,安徽 淮南 232001
2. 安徽省煤矿安全大数据分析和预警技术工程实验室,安徽 淮南 232001

摘 要:业务流程中的一项重要工作是进行数据的异常检测,它可以用于监控和识别企业或组织中出现的异常情况。**目的** 针对目前业务流程异常检测方法大多数只考虑控制流,并未考虑事件日志中其他数据属性对业务流程影响的情况,提出一个多视角无监督异常检测模型。**方法** 首先,将控制流和数据流分别进行处理,然后拼接形成可以输入到模型中的数据类型;其次,利用自注意力机制和 Bi-LSTM 自编码器组合成的模型,分别对控制流视角和数据流视角进行业务流程事件日志的特征提取,并进行拼接和异常检测,异常阈值由自编码器的重构误差来确定;最后将提出的模型在公共数据集上进行了验证。**结果** 用真实事件日志对提出的方法进行评估,与其他方法进行对比分析可知,所提出的方法在精确度、召回率和 F1 分数 3 个方面都有较好的表现,且所提出的模型 AUC 在所有数据集上都达到了较大的值。**结论** 实验结果表明:所提出的方法可以更好地检测过程事件日志中的异常;通过在模型中加入注意力机制并且将控制流和数据流视角进行结合,更好地表示了过程数据,使得模型的分类性能得到了较大的提升,在业务流程异常检测方面具有明显的优势。

关键词:自注意力机制;Bi-LSTM 神经网络;业务流程;异常检测
中图分类号:TP391.9 **文献标识码**:A doi:10.16055/j.issn.1672-058X.2025.0002.015

Anomaly Detection Model for Business Process Based on Self-attention and Bi-LSTM

CHEN Guowei[1], LU Ke[1,2]
1. School of Mathematics and Big Data, Anhui University of Science and Technology, Anhui Huainan 232001, China
2. Anhui Province Engineering Laboratory for Big Data Analysis and Early Warning Technology of Coal Mine Safety, Anhui Huainan 232001, China

Abstract: An important task in business processes is anomaly detection of business process data, which can be used to monitor and identify abnormal situations in enterprises or organizations. **Objective** Most current methods for business process anomaly detection only consider the control flow and do not consider the influence of other data attributes in event logs on business processes. Therefore, an unsupervised anomaly detection model with multiple perspectives was proposed. **Methods** Firstly, the control flow and data flow were processed separately and then spliced to form the input data type for the model. Secondly, a model combining the self-attention mechanism and Bi-LSTM autoencoder was used to extract features from the perspectives of control flow and data flow of business process event logs respectively, and then splicing was carried out for anomaly detection, with the anomaly threshold determined by the reconstruction error of the autoencoder. Finally, the proposed model was validated on public datasets. **Results** The proposed method was evaluated using real event logs, and a comparative analysis with other methods showed that the proposed method performed better in three aspects: precision, recall, and F1 score, and the AUC of the proposed model reached large values on all datasets.

收稿日期:2023-10-24 **修回日期**:2023-12-14 **文章编号**:1672-058X(2025)02-0112-08
基金项目:国家自然科学基金资助项目(61402011);安徽省重点研究与开发计划项目(2022A05020005).
作者简介:陈国威(1998—),男,安徽宿州人,硕士研究生,从事过程挖掘研究.
通讯作者:卢可(1995—),男,安徽宿州人,博士研究生,从事过程挖掘、业务流程监控研究. Email:kelu@aust.edu.cn.
引用格式:陈国威,卢可. 基于自注意力和 Bi-LSTM 的业务流程异常检测模型[J]. 重庆工商大学学报(自然科学版),2025,42(2):112-119.
CHEN Guowei, LU Ke. Anomaly detection model for business process based on self-attention and Bi-LSTM[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2025, 42(2): 112-119.

Conclusion Experimental results show that the proposed method can better detect anomalies in process event logs. By incorporating attention mechanisms into the model and combining control flow and data flow perspectives, a better representation of process data is achieved, leading to significantly improved classification performance and clear advantages in business process anomaly detection.

Keywords: self-attention mechanism; Bi-LSTM neural network; business process; anomaly detection

1 引　言

由于流程的参与者众多,流程会变得特别复杂,同时信息系统(Information System,IS)也被纳入业务流程中,并产生越来越多的日志数据,这给流程管理和数据分析带来了全新的挑战。过程感知信息系统[1](Process Aware Information System,PAIS)已经在各个领域中用于处理流程业务。在业务流程运行过程中,可能出现可预测的异常,如资源异常或流程时间异常,而有些则是无法预测的。异常情况对业务流程的影响非常大,可能会给用户带来巨大的损失,例如在进行贷款时对未经审查征信信息的申请者进行放贷。依靠过程感知信息系统(PAIS)会产生相应的事件日志,这些事件日志记录了流程中的活动。因此,可以对这些事件日志进行流程分析。过程挖掘(Process Mining,PM)是一个从过程数据(事件日志)中提取信息和进行过程分析的领域。过程挖掘使信息系统能够对各种异常情况和偏差做出反应。检测异常流程实例是企业组织特别关注的一方面,因为检测不正确的执行可以避免欺诈并节省资源[2-3],从而将损失最小化或避免损失的发生。异常检测方法在业务流程数据分析领域越来越受欢迎,对于实时决策的需求也越来越大[4]。

文献[5]提出了一种使用自编码器进行异常检测的方法,该方法描述了允许在没有先验知识的情况下进行事件日志异常识别,但只使用了控制流视单一视角;文献[6]介绍了一种用于多变量业务流程异常检测的深度学习方法,其中,描述了一种用于事件日志中实时多变量异常检测的网络架构;基于文献[6]的这项工作,该作者在文献[7]中描述了业务流程异常的多视角分类;文献[8]提出了一个分类模型,该模型考虑了深度学习、时间序列分析和序列处理,并将它们组合成一种方法,该方法能够从事件数据中过滤出与活动和时间相关的异常。

文献[9]通过使用流程事件日志数据的图形编码和图形自动编码器构建异常检测器来识别业务流程中的异常,该方法不需要业务流程先验知识;文献[10]提出一种无监督在线流程挖掘环境中应用变分自动编码器过滤异常事件数据的方法,该方法有助于改进过程挖掘技术的结果,从而对业务流程管理产生积极影响;文献[11]提出一种方法并定义了一种称为基于深度学习编码器的异常检测(DLE-AD)算法,增强了异常检测的能力;文献[12]考虑在嘈杂的业务流程事件日志情况下,使用无监督方法进行异常检测,他们采用 BPMN 模型对流程进行可视化表示,并使用基于深度学习的自动编码器来检测异常;文献[13]提出一种基于上下文感知的多角度业务流程在线异常检测方法,该方法对当前执行实例进行在线检测;文献[14]提出一种基于机器学习方法的异常检测方法,该方法可以实时预测业务流程中的超期异常和流程行为异常。

上述研究成果对业务流程异常检测都具有重大意义,但其中部分方法仅仅使用控制流视角进行研究,无法将数据流视角考虑进去。部分方法虽考虑了多视角(控制流+数据流),但在精度及 F1 分数上表现不太理想,没有明显的提升,也没有对业务流程各属性重要性进行考虑。因此本文提出了一种基于深度学习的分类方法,即使用自注意力机制和自编码器(Autoencoder,AE)进行业务流程数据的异常检测。该方法考虑了控制流视角和数据流视角,基于无监督学习方法,使用基于重构的方法从业务流程数据中检测异常,同时使用双向长短期记忆网络(Bi-directional Long Short-Term Memory,Bi-LSTM)来构建自编码器,使得自编码器可以在分析过程中更充分地捕捉到事件序列数据中的长期依赖关系。

此外,自注意力机制可以进一步捕捉输入数据中的长期依赖关系,该机制可以学习不同位置之间的依赖关系,并计算权重,从而更好地捕捉输入数据中的重要特征。自注意力机制的引入,可以提高自编码器模型的泛化能力以更好地适应新的数据,同时也能够提高模型的计算效率。

2 预备知识

2.1 过程挖掘

过程挖掘主要关注的是如何从业务流程中的数据

(事件日志)中提取出有价值的信息,以帮助企业更好地理解、优化和管理业务过程。在业务流程执行过程中,每个流程的步骤都会被记录在数据库中。包括什么时候执行了流程的步骤(时间戳),执行了哪一个流程步骤(活动),由谁执行了这个步骤(资源),以及这个步骤属于哪一个业务流程案例(案例标识符)。这些事件信息是过程挖掘算法的基础。所有的流程案例组合成一个被称为事件日志的数据结构。

事件日志是由过程感知信息系统 PAIS 记录业务流程执行过程中信息的集合,它是由多个案例组合而成的。每个案例都是一个业务流程执行实例,它由流程中的一系列事件组成。每个事件都由活动名称及其属性组成。下面将介绍相关定义。

定义 1 事件是业务流程实例中发生的最小单元,包含活动和其他相关信息;属性描述事件涉及的活动特征[15]。

一般的事件属性包括事件的活动、资源、时间戳等,这些属性是每个事件的基本属性。

定义 2 案例由业务流程中执行事件的序列组成,表示流程执行的迹,用 $e=(case\_id, event\_id, attr_{\text{activity}}, attr_{\text{time}}, \cdots, attr\cdots)$ 表示。$case\_id$ 表示事件的案例 id,$event\_id$ 表示事件 id,$attr_{\text{activity}}$ 表示事件的活动名称,$attr_{\text{time}}$ 表示事件的时间戳,$attr\cdots$ 表示事件的其他相关属性。

定义 3 C 是所有案例的集合,E 是所有事件的集合,A 是所有事件属性的集合,事件日志 L 是案例 $L\subseteq C$ 的集合,事件日志可以表示为 $L=R^{|C|\times|E|\times|A|}$。事件日志示例如表 1 所示。

表 1 事件日志示例

Table 1 Example of an event log

ID	时间戳	活 动	资 源
1	2015-03-21 12:38:39	PR created	Roy
1	2015-03-28 07:09:26	PR released	Earl
1	2015-04-21 16:59:49	Goods receipt	Ryan
2	2015-05-14 11:31:53	SC created	Marilyn
2	2015-05-21 09:21:26	SC purchased	Emily
2	2015-05-28 18:48:27	SC approved	Roy

案例异常指业务流程执行过程中出现的不合理行为或属性。在事件日志 L 中,绝大部分的流程案例都可以正常运行直到流程完成,只有少数的流程案例会发生异常。这些异常在事件日志中通常表现为离群的事件或属性值。本文的研究内容主要是进行业务流程运行案例的异常检测。

2.2 Bi-LSTM

长短期记忆网络(LSTM)是 RNN 的变体,通过门控机制结合了短期和长期记忆,以解决 RNN 中的梯度消失问题,实现对长序列依赖性的建模。LSTM 单元结构如图 1 所示,LSTM 中有 3 个门。

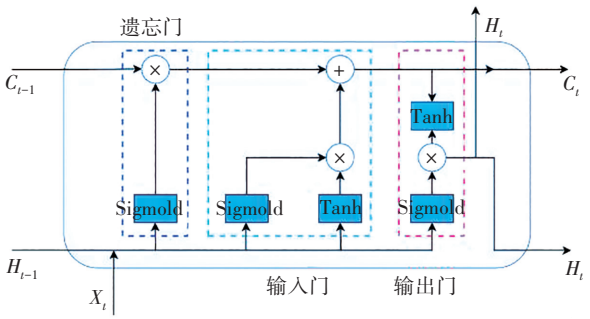


图 1 一个 LSTM 单元

Fig. 1 An LSTM cell

遗忘门:LSTM 网络中的遗忘门决定哪些隐藏信息应该被遗忘或丢弃。它利用 Sigmoid 函数来判断哪些细节需要从记忆中删除。

$$f_t=\text{Sigmoid}(\boldsymbol{W}_f\times[\boldsymbol{H}_{t-1},\boldsymbol{X}_t]+b_f) \tag{1}$$

其中,f_t 是遗忘门的输出,$\boldsymbol{W}_f$、$\boldsymbol{H}_{t-1}$、$\boldsymbol{X}_t$ 分别是权重矩阵、前一个隐藏状态和当前输入,b_f 是偏置项。

输入门:输入门可以选择性地将当前时刻的输入信息加入记忆状态中。

$$i_t=\text{Sigmoid}(\boldsymbol{W}_i\times[\boldsymbol{H}_{t-1},\boldsymbol{X}_t]+b_i) \tag{2}$$

$$\hat{\boldsymbol{C}}_t=\text{Tanh}(\boldsymbol{W}_c\times[\boldsymbol{H}_{t-1},\boldsymbol{X}_t]+b_c) \tag{3}$$

$$\boldsymbol{C}_t=f_t\times\boldsymbol{C}_{t-1}+i_t\times\hat{\boldsymbol{C}}_t \tag{4}$$

其中,$\boldsymbol{C}_{t-1}$ 是旧的单元状态,Sigmoid 函数决定哪些值需要更新;Tanh 会生成一个值$\hat{\boldsymbol{C}}_t$,这个值将会被添加到新的单元状态中;$\boldsymbol{C}_t$ 是传递给下一步的新单元状态。

输出门:输出门负责生成新的隐藏状态以及时间步长的输出,结合输入和记忆来确定输出。

$$o_t=\text{Sigmoid}(\boldsymbol{W}_o\times[\boldsymbol{H}_{t-1},\boldsymbol{X}_t]+b_o) \tag{5}$$

$$\boldsymbol{H}_t=o_t\times\text{Tanh}(\boldsymbol{C}_t) \tag{6}$$

其中,$\boldsymbol{H}_t$ 是新的隐藏状态。

双向长短期记忆网络(Bi-LSTM)和 LSTM 的原理相同,是 LSTM 的一种变体,由前向 LSTM 与后向 LSTM

组合而成。LSTM 通过训练可以学习记忆和遗忘的能力。但是 LSTM 只能进行前向传播编码,无法捕获反向上下文信息。通过 Bi-LSTM 就可以进行双向的序列信息捕捉,能更好地将信息进行编码。

$$\overrightarrow{\boldsymbol{h}_i}=\overrightarrow{\mathrm{LSTM}}(a_{i,j},h_{i,j-1}),j=1,2,\cdots,m \tag{7}$$

$$\overleftarrow{\boldsymbol{h}_i}=\overleftarrow{\mathrm{LSTM}}(a_{i,j},h_{i,j+1}),j=1,2,\cdots,m \tag{8}$$

$$\boldsymbol{h}_i=[\overrightarrow{\boldsymbol{h}_i},\overleftarrow{\boldsymbol{h}_i}] \tag{9}$$

其中,$a_{i,j}$ 代表事件 i 中的第 j 个属性,h 是 LSTM 的隐藏状态。$\overrightarrow{\boldsymbol{h}_i}$ 是前向隐藏状态,$\overleftarrow{\boldsymbol{h}_i}$ 为后向隐藏状态,$\overrightarrow{\mathrm{LSTM}}$ 是前向 LSTM,$\overleftarrow{\mathrm{LSTM}}$ 是后向 LSTM。$\boldsymbol{h}_i$ 总结了事件 i 的邻居事件,同时仍然关注着事件 i。Bi-LSTM 的架构如图 2 所示。

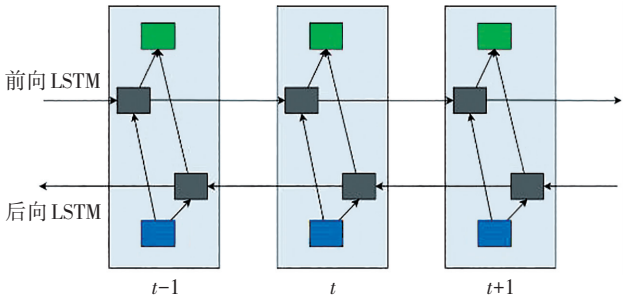


图 2　Bi-LSTM 框架

Fig. 2　Bi-LSTM framework

2.3　自注意力机制

自注意力机制[16]是注意力机制的变体,它减少了对外部信息的依赖,能更好地捕捉数据或特征之间的内部关系,并且能同时考虑序列中的所有元素。自注意力机制通过计算每个元素与其他元素之间的相似度来得到每个元素的权重,然后将所有元素加权求和得到最终的加权表示向量,来表示当前位置输入与其他位置输入的关联性。自注意力机制的核心是捕捉向量之间的相关性。自注意力机制 S_A 可以被看作是查询(Q)和键($\boldsymbol{K}$)-值($\boldsymbol{V}$)对之间的一个映射模型。可以表示为

$$S_A(\boldsymbol{Y})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_K}}\right)\boldsymbol{V}$$

在式(10)中,$\boldsymbol{Q}$、$\boldsymbol{K}$ 以及 $\boldsymbol{V}$ 是通过线性变换得到的:

$$\boldsymbol{Q}=\boldsymbol{W}_q\boldsymbol{Y}\ ,\ \boldsymbol{K}=\boldsymbol{W}_k\boldsymbol{Y}\ ,\ \boldsymbol{V}=\boldsymbol{W}_v\boldsymbol{Y} \tag{11}$$

其中,$\boldsymbol{W}_q$、$\boldsymbol{W}_k$ 以及 $\boldsymbol{W}_v$ 是权重矩阵 $\boldsymbol{Q}$ 和 $\boldsymbol{K}^{\mathrm{T}}$ 的乘积,用来计算权重矩阵;Softmax 函数被用来对权重进行归一化到[0,1]之间,同时为了避免注意力矩阵值过大,利用缩放点积模型来调整权重值;点积结果除以 $\boldsymbol{K}$ 的特征维度的平方值,最后权重矩阵和 $\boldsymbol{V}$ 的乘积就是当前节点新的特征向量。自注意力机制的框架如图 3 所示。

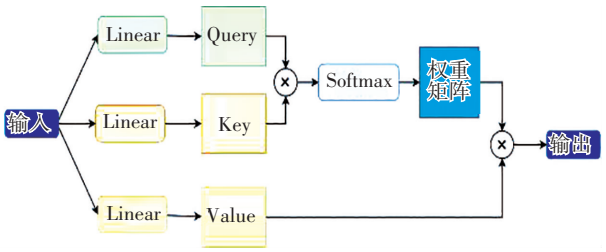


图 3　自注意力机制

Fig. 3　Self-attention mechanism

3　模型设计与检测方法

在这一节中,将描述本文所提出的业务流程异常检测方法。正如先前所描述,每个事件除了活动名称可能有许多属性,如资源、时间戳等。这些属性在进行异常检测时可能是有参考价值的,因此,也应当被考虑进去。

3.1　数据预处理

在事件数据可用于训练和测试分类模型之前,必须首先使其适合机器学习算法的处理。预处理需要执行 5 个基本步骤。事件数据的转换包括以下步骤:

数据集统计数据的测量:包括使用统计数据识别值中的总体异常值,例如查看周期时间、识别重复项以及标记数据集中的缺失值。

缺失值的估算:包括用基于现有信息的估计数据值替换所考虑数据集中缺失或不一致的数据值。

检查不一致:在对数据集进行初步清理和修正后,重新检查数据集是否存在严重的不一致。

对轨迹进行编码:尽可能高效地对轨迹进行编码,可以获得更好的分类结果。

数据集标准化:数据集的归一化用于统一数值范围不同的数字数据值的比例,而不会扭曲数值范围的差异。

在对数据集进行处理和转换的基础上,得到预处理后的数据集,可将其传递给分类模型进行训练和测试。

3.2　模型设计

本文提出了一个基于自注意力机制的自编码器进行异常检测的模型,可以自动学习在异常检测中需要更加关注的信息。图 4 为模型的结构。

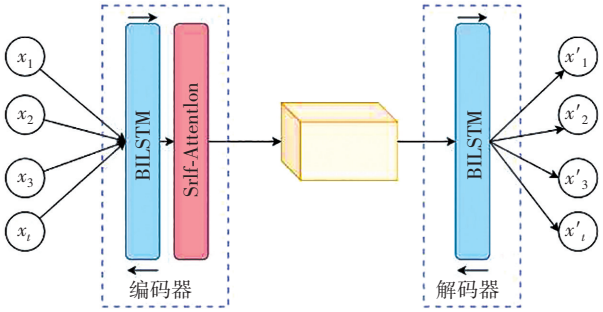


图 4 基于自注意力机制的自编码器结构

Fig. 4 Self-encoder structure based on self-attention mechanism

自编码器模型接受一连串的事件数据 $x=(x_1,x_2,x_3,\cdots,x_t)$ 作为输入,这个输入序列被传递给编码器。如前所述,本文使用 Bi-LSTM 作为自编码器的基础,Bi-LSTM 编码器可以根据输入序列生成正向和反向的隐藏状态序列,隐藏状态构成一个可以用来存储过去和已知信息的模型组件,因此可以用于需要考虑长期依赖关系的任务。例如本文中所进行的业务流程异常检测。通过使用 Bi-LSTM,可以从两个方向来查看输入序列,进而可以识别输入序列的结构关系。编码器生成的隐藏状态 $\boldsymbol{h}_t$ 组成一个向量,并将向量传递到自注意力网络层中。

自注意力层的输入是从编码器中传入的一系列隐藏向量,并由此给出一系列相应的上下文向量 $\boldsymbol{c}_t$,这些向量是所有输入向量的加权和。使用缩放点积[16]确定编码器的成对隐藏状态 $\boldsymbol{h}_i$ 和 $\boldsymbol{h}_j$ 的相关性,其维度是 d_h,对得到的分数使用 Softmax 函数确定向量中值的权重。

自注意力机制将上下文信息从编码器的隐藏状态经过加权后传递给解码器。从加权编码器的隐藏状态中,确定所有的上下文向量 $\boldsymbol{c}_t$ 后,将向量传递给解码器进行处理,解码器根据上下文向量 $\boldsymbol{c}_t$ 的信息确定哪些隐藏状态 $\boldsymbol{h}_t$ 需要单独进行考虑。

潜在表征 $\boldsymbol{z}$ 和上下文向量 $\boldsymbol{c}_t$ 一起被传递给 Bi-LSTM 解码器,在每个时间段 t 内,和 Bi-LSTM 编码器的情况类似,首先生成隐藏状态 $\boldsymbol{h}_t$,然后通过 Bi-LSTM 解码器确定隐藏状态 $\boldsymbol{h}_t$,接着利用隐藏状态生成网络的输出。输出的是输入层输入 x_t 的重构序列 x'_t,隐藏状态被全连接层连接,确保隐藏状态 $\boldsymbol{h}_t$ 可以被重构为 x'_t。图 4 说明了所提出的模型,通过结合自注意力机制和自编码器的方法,可以将这个模型用于业务流程事件数据的异常检测。

3.3 异常检测

目前,基于重构的数据异常检测方法被认为可以尽可能地减少信息损失,将输入数据进行重构,重构过程中可能会出现数据损失,这种损失被称为重构误差。本文模型首先用正常数据进行训练,以学习正常数据的特征行为。通常正常数据比异常数据能够更好地进行重构。因此,模型可以利用这个方法来区分正常数据和异常数据。将输入数据和重构数据之间的距离定义为重构误差。

$$\tau=\frac{1}{n}\sum_{i=1}^{n}e_i^{\,2} \tag{12}$$

$$\sigma(s,\tau)=\begin{cases}1, s>\tau\\ 0, \text{否则}\end{cases} \tag{13}$$

重构误差被用作计算异常分数,并计算出一个阈值,超过这个阈值的数据被标记为异常。阈值用 τ 表示,e_i 是迹的重构误差,n 是迹的数量。如果重构误差大于设定阈值,会被识别为异常,反之为正常。

4 仿真实验与结果分析

4.1 数据集

本文采用 BPIC(Business Process Intelligence Challenge)中的 BPIC12、BPIC13 和 BPIC17 事件日志作为数据集,这些数据集的具体信息如表 2 所示。

表 2 数据集特征

Table 2 Characteristics of the dataset

数据	日志	活动	案例	事件	属性
BPIC12	1	73	13 k	290 k	0
BPIC13	3	27	0.8 k~7.5 k	4 k~81 k	4
BPIC17	3	53	31 k~43 k	284 k~1.2 m	1

由于 BPIC 的原始事件日志中包含小部分异常,在现实中很难对这些异常进行区分和标记,因此可以通过进行人工插入异常的方法构建数据集,这种方法可能会增加异常检测的难度,但也可以提高算法的鲁棒性和泛化能力。为了尽可能模拟现实生活中的异常,本文在数据集中插入 6 种类型的异常。随机从事件日志中挑选 30%的案例进行异常插入。引入了 Original 表示 Early 或 Late 活动所在的原始位置,6 种异常类型如下所述:

- Skip:跳过 n 个事件。
- Insert:随机插入 n 个随机活动。
- Rework:随机选择位置插入案例的 n 个事件。
- Early:提前案例的 n 个事件。
- Late:将案例的 n 个事件推迟。
- Attribute:将案例的 n 个事件属性设置不正确的值。

本文取 $n\in(2,3)$ 来进行实验,异常类型如图 5 所示。

图 5　异常示例

Fig. 5　Example of an exception

4.2　结果分析

对于上述数据集,本文进行了基于控制流视角和数据流视角组合的异常检测,并将其与 Sampling[17]、DAE[5]、BINet[6]、Likelihood[18] 和 OS - SVM[19] 进行比较。

为了验证本文所提出方法在多视角业务流程异常检测方面的有效性,采用 F1-score 作为综合指标来进行评估。F1-score 是一种结合精确率和召回率的评价指标,可以全面反映出模型的异常检测性能,最大值是 1,最小值是 0,值越大表示模型越好。精确度(Precision)和召回率(Recall)是衡量二分类模型性能的两个重要指标。精确度指的是在所有被模型预测为正例的样本中,有多少是真正的正例。召回率指的是在所有真正的正例中,有多少被模型正确地预测为正例。

在本文中 F1-score 越高,表明模型的异常检测效果越好,本文的工作只聚焦于案例级的异常检测。表 3 显示在数据集 BPIC12 和 BPIC13 上本文的方法在 F1-score 上优于现有方法,在 BPIC17 上优于大部分方法,达到了平均水平。本文的实验评估结果如表 3 所示。

表 3　不同方法的 F1-score 对比

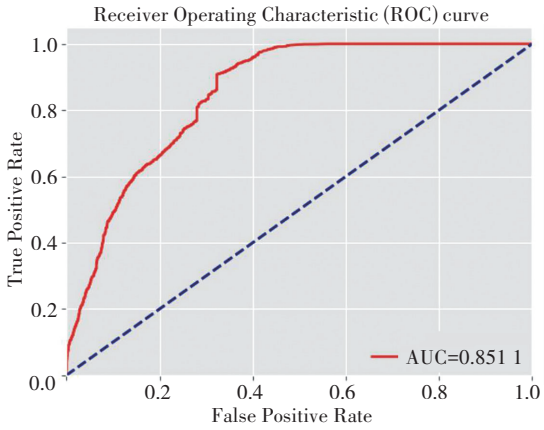
Table 3　Comparison of F1-score of different methods

方法	数据集		
	BPIC12	BPIC13	BPIC17
Sampling	0. 55	0. 21	0. 32
DAE	0. 60	0. 34	0. 60
BINet	0. 61	0. 41	0. 47
Likelihood	0. 62	0. 45	0. 40
OS-SVM	0. 55	0. 25	0. 35
SA-BiLSTM-AE(本文)	0. 75	0. 45	0. 57

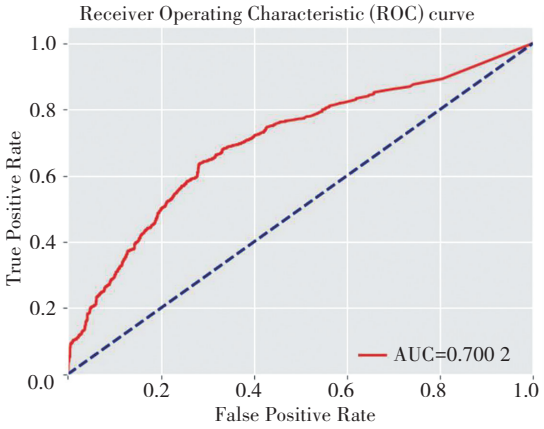
对于 BPIC12、BPIC13 和 BPIC17 这 3 个数据集,数据集 BPIC12 和 BPIC17 所对应的过程模型相对简单,因此所有方法在这两个数据集上的 F1-score 比较高。而 BPIC13 数据集的过程模型相对复杂,所有方法在这个数据集上的表现不如另外两个数据集。对于不同数据集的复杂度,模型越简单,分类效果越好。可以看出,随着过程模型复杂性的增加,所提出模型的分类质量略有下降。过程数据越复杂的结构,对于分类模型来说更难以进行泛化。

ROC(Receiver Operating Characteristic)曲线是一种用于评估分类器性能的常用方法。在二元分类问题中,TPR 也被称为灵敏度(Sensitivity),即真正类被正确预测为正类的比例,FPR 则被称为 1-特异度(1-Specificity),即真负类被错误预测为正类的比例。AUC(Area Under Curve),即 ROC 曲线下的面积,反映了所提出方法区分异常数据数目和正常数据数目的能力,AUC 值越大,分类器的性能越好。

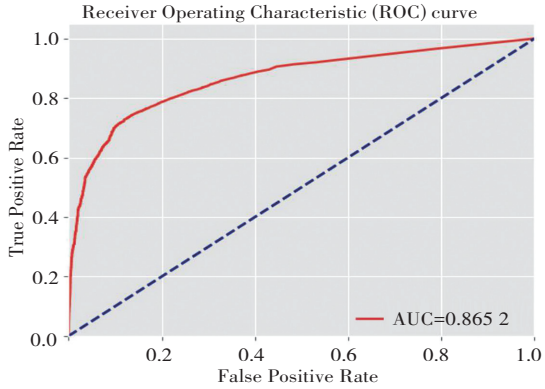
图 6 展示了本文方法在 3 个数据集上的 ROC 曲线,该方法在 BPIC12 数据集上的 AUC 达到了 0. 851 1,在 BPIC13 数据集上的 AUC 达到了 0. 700 2,在 BPIC17 数据集上的 AUC 达到了 0. 865 2。结果反映了本文所提出方法在异常检测方面具有良好的分类能力。



(a) BPIC12 的 ROC 曲线



(b) BPIC13 的 ROC 曲线

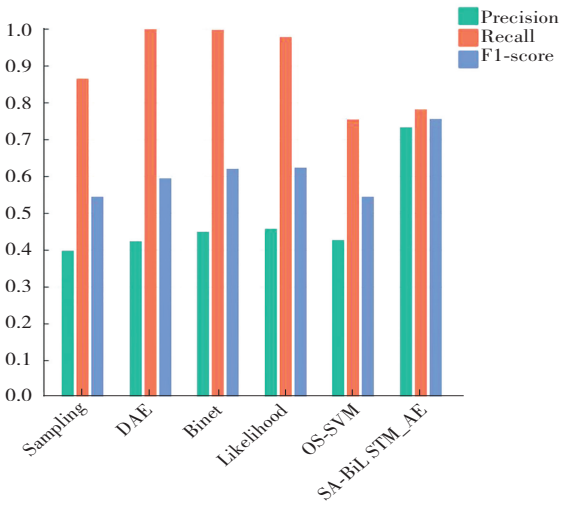


(c) BPIC17 的 ROC 曲线

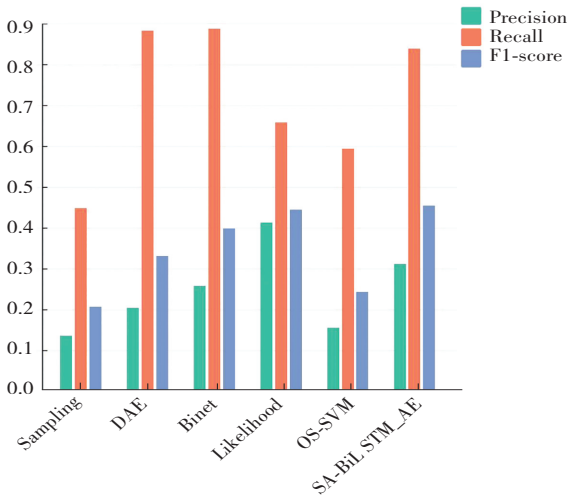
图 6 本文方法在 3 个数据集上的 ROC 曲线

Fig. 6 ROC curves of the proposed method on three datasets

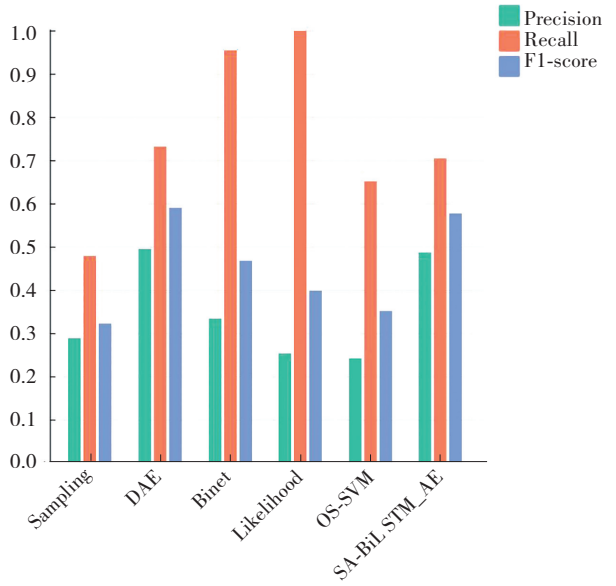
图 7 展示了 BPIC12、BPIC13 以及 BPIC17 3 个数据集在精确度、召回率 和 F1-score 上和选择的比较方法的对比。从图 7 中综合来看,本文的方法在精确度、召回率和 F1-score 3 个评价指标上,优于大部分方法。从上述结果可以得出,本文的方法可以有效地检测出业务流程案例级的异常。



(a) BPIC12



(b) BPIC13



(c) BPIC17

图 7 各种方法在 3 个数据集上的精确度、召回率和 F1-score 值对比图

Fig. 7 Comparison chart of precision, recall and F1-score of various methods on three datasets

通过对实验结果图 6、图 7 的分析,可以看出本文提出的基于多视角的异常检测方法可以有效地将业务流程案例级的异常检测出来,并且在精确度方面优于大多数方法,在召回率方面达到了所有比较方法的平均水平,对于 F1-score 来说,本文的方法优于大多数比较方法。

5 结 论

提出了一种基于深度学习方法的无监督异常检测分类模型,该模型可以用于识别业务流程事件序列中的异常。通过使用双向 LSTM 递归神经网络作为该模型的编码器和解码器,可以考虑事件数据中的时间依赖性。通过集成自注意层,该模型能够更好地处理来自编码器隐藏状态的某些信息。实验评估部分表明:该方法可以在事件日志数据集中产生比较好的结果,与其他工作相比,本文的方法在 BPIC12、BPIC13 以及 BPIC17 3 个数据集上的分类结果优于现有方法的平均水平。

未来的工作将专注于事件日志属性级别的异常检测,同时尝试将这种方法应用于过程感知信息系统的在线过程异常检测,并探索在业务流程中预测下一活动和剩余时间的应用。

参考文献(References):

[1] VANDER AALST W. Process mining: Data science in action [M]. 2nd ed. Heidelberg: Springer, 2016.

[2] JAGADEESH CHANDRA BOSE R, VAN DER AALST W. Trace alignment in process mining: opportunities for process diagnostics[C]//Lecture Notes in Computer Science. Heidelberg: Springer Berlin Heidelberg, 2010: 227-242.

[3] TAVARES G M, BARBON S Jr. Analysis of language inspired trace representation for anomaly detection [C]// Communications in Computer and information Science. Cham: Springer International Publishing, 2020: 296-308.

[4] KRAJSIC P, FRANCZYK B. Lambda architecture for anomaly detection in online process mining using autoencoders [C]// Communications in Computer and information Science. Cham. Springer International Publishing, 2020: 579-589.

[5] NOLLE T, LUETTGEN S, SEELIGER A, et al. Analyzing business process anomalies using autoencoders [J]. Machine Learning, 2018, 107(11): 1875-1893.

[6] NOLLE T, SEELIGER A, MÜHLHÄUSER M. BINet: Multivariate business process anomaly detection using deep learning [C]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018: 271-287.

[7] NOLLE T, LUETTGEN S, SEELIGER A, et al. Binet: Multi-perspective business process anomaly classification[J]. Information Systems, 2022, 103(10): 1-12.

[8] KRAJSIC P, FRANCZYK B. Semi-supervised anomaly detection in business process event data using self-attention based classification[J]. Procedia Computer Science, 2021, 192(8): 39-48.

[9] HUO S, VÖLZER H, REDDY P, et al. Graph autoencoders for business process anomaly detection[C]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021: 417-433.

[10] KRAJSIC P, FRANCZYK B. Variational autoencoder for anomaly detection in event data in online process mining[C]// Proceedings of the 23rd International Conference on Enterprise Information Systems. SciTePress: Science and Technology Publications, 2021: 567-574.

[11] VIJAYAKAMAL M, VASUMATHI D. A novel approach to detect anomalies in business process event logs using deep learning algorithm[C]//Advances in intelligent Systems and Computing. Singapore: Springer Singapore, 2022: 363-374.

[12] NOLLE T, SEELIGER A, MÜHLHÄUSER M. Unsupervised anomaly detection in noisy business process event logs using denoising autoencoders[C]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016: 442-456.

[13] 孙笑笑, 侯文杰, 沈沪军, 等. 基于上下文感知的多角度业务流程在线异常检测方法[J]. 计算机集成制造系统, 2021, 27(9): 2532-2541.
SUN Xiao-xiao, HOU Wen-jie, SHEN Hu-jun et al. Context-aware online anomaly detection method for multi-angle business processes[J]. Computer-integrated Manufacturing Systems, 2021, 27(9): 2532-2541.

[14] 魏懿, 曹健. 基于机器学习的流程异常预测方法[J]. 计算机集成制造系统, 2019, 25(4): 864-872.
Wei Yi, Cao Jian. A machine learning based process anomaly prediction method[J]. Computer Integrated Manufacturing Systems, 2019, 25(4): 864-872.

[15] RUSSELL N, VAN DER AALST W, TER HOFSTEDE A. Workflow exception patterns[C]//Notes on Numerical Fluid Mechanics and Multidisciplinary Design. Cham: Springer international Publishing, 2006: 288-302.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc, 2017: 6000-6010.

[17] BEZERRA F, WAINER J. Algorithms for anomaly detection of traces in logs of process aware information systems[J]. Information Systems, 2013, 38 (1): 33-44.

[18] BÖHMER K, RINDERLE-MA S. Multi-perspective anomaly detection in business process execution events[C]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016: 80-98.

[19] WARRENDER C, FORREST S, PEARLMUTTER B. Detecting intrusions using system calls: Alternative data models[C]//Proceedings of the 1999 IEEE symposium on Security and Privacy. Piscataway: IEEE Press. 1999: 133-145.

责任编辑:李翠薇