基于改进 BERT 和轻量化 CNN 的业务流程合规性检查方法

田银花[1]，杨立飞[1]，韩咚[2*]，杜玉越[3]

（1. 山东科技大学智能装备学院，山东 泰安 271000；2. 山东科技大学继续教育学院，山东 泰安 271000；
3. 山东科技大学计算机科学与工程学院，山东 青岛 266590）

摘　要：业务流程合规性检查可以帮助企业及早发现潜在问题，保证业务流程的正常运行和安全性。提出一种基于改进 BERT（Bidirectional Encoder Representations from Transformers）和轻量化卷积神经网络（CNN）的业务流程合规性检查方法。首先，根据历史事件日志中的轨迹提取轨迹前缀，构造带拟合情况标记的数据集；其次，使用融合相对上下文关系的 BERT 模型完成轨迹特征向量的表示，最后，使用轻量化 CNN 模型构建合规性检查分类器，完成在线业务流程合规性检查，有效提高合规性检查的准确率。在 5 个真实事件日志数据集上进行实验，结果表明，该方法相比 Word2Vec＋CNN 模型、Transformer 模型、BERT 分类模型在准确率方面有较大提升，且与传统 BERT＋CNN 相比，所提方法的准确率最高可提升 2.61%。

关键词：业务流程；合规性检查；表示学习；事件日志；卷积神经网络

源代码链接：https://github.com/flagfly6/CtxtBERT

中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0069357

Conformance Checking Method of Business Processes Based on Improved BERT and Lightweight CNN

TIAN Yinhua[1], YANG Lifei[1], HAN Dong[2*], DU Yuyue[3]

(1. College of Intelligent Equipment, Shandong University of Science and Technology, Tai'an 271000, Shandong, China;
2. College of Continuing Education, Shandong University of Science and Technology, Tai'an 271000, Shandong, China;
3. College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, Shandong, China)

【Abstract】 Business process conformance checking can help enterprises detect potential problems early and ensure the normal operation and security of business processes. To this end, a conformance checking method for business processes based on improved Bidirectional Encoder Representations from Transformers (BERT) and lightweight Convolutional Neural Network (CNN) is proposed. First, trace prefixes are extracted from historical event logs and labeled with fitness or unfitness, and a dataset is constructed accordingly. Second, the improved BERT model is used to represent the feature vectors of traces, which incorporates relative contextual relationships. Finally, the conformance check classifier, constructed using a lightweight CNN model, is used to complete online business process conformance checking. This method effectively improves the accuracy of conformance checking. Experiments were conducted using five real-life event log datasets. The results show that the proposed model's accuracy is superior to that of Word2Vec＋CNN, Transformer, BERT. Furthermore, when compared with the traditional BERT＋CNN, the accuracy can increase by up to 2.61%.

【Key words】 business process; conformance checking; representation learning; event log; Convolutional Neural Network (CNN)

0　引言

大数据技术的不断发展极大地促进了业务过程管理的提升[1-2]。在激烈的市场竞争中，企业需要更加智能、高效、有序地进行业务过程管理，提高生产效率，从而满足不断变化的市场需求[3]。大数据的发展不仅提供了大量、多样化的数据，而且提供了存储和分析这些海量数据的能力，使得企业可以记录、分析这些多样化的数据，并且能够更好地理解和管理自己的业务过程[4-5]。通过大数据的分析，企业不仅可以发现隐藏在数据背后的关联关系，挖掘潜在客户，而且可以实时监控和预测业务过程的变化，迅速发现不合规的情况，从而采取相应措施，最大程度地保证业务流程的稳定。

收稿日期：2024-02-05　　**修回日期：**2024-04-08

基金项目：国家自然科学基金（72101137，61973180）；教育部人文社会科学研究青年基金项目（21YJCZH150，20YJCZH159）；山东省自然科学基金（ZR2021MF117，ZR2022QF020）；山东省重点研发计划（软科学）项目（2022RKY02009）；山东省习近平新时代中国特色社会主义思想研究中心山东科技大学山东数字经济研究基地项目（SDSZJD202314）。

通信作者 E-mail：* aa1130_2011@163.com

不合规事件的发生可能会对企业的效率、质量和利润产生负面影响[6]。因此,企业的业务流程合规性检查变得至关重要。通过建立有效的业务流程合规性检查方法,能够实时监测企业运营过程中的不合规情况[7],帮助企业及时发现问题并解决潜在危机,使得企业能够减少损失,提升客户满意度,增强竞争力。

合规性检查包括离线合规性检查和在线合规性检查。离线合规性检查是指对已经收集到的历史数据进行合规性检查和识别,这种方法通常使用过程挖掘等技术识别不合规情况[8]。过程挖掘在业务流程管理中的作用是巨大的,对于业务过程的合规性检查研究颇有意义。通过建立业务过程的模型,企业可以监控和识别不合规活动和行为,及时发现并处理潜在问题[9]。当前大多数的合规性检查技术都是离线工作的,即对已经执行完成的轨迹进行合规性检查,但是该类方法缺乏实时性,不能立刻洞察出错行为[10]。在线合规性检查是指在实时数据流中进行合规性检查和识别,其主要特点是能够实时监测数据流并及时检测不合规情况,最大程度地帮助企业减少潜在的损失[11]。本文的主要贡献如下:

1)对 BERT(Bidirectional Encoder Representations from Transformers)模型进行改进优化。为了充分利用业务流程事件日志中整条轨迹的上下文信息,将 BERT 模型与相对上下文关系进行融合。通过在预训练过程中学习上下文关系,更准确地捕捉轨迹中单个活动之间的关联性;对业务流程中因循环结构或选择结构产生的相同活动序列片段,能够更好地学习特征向量表示。

2)针对业务流程的合规性检查,提出一种通用的基于表示学习的业务流程合规性检查框架,用于实现不同长度事件日志的业务流程合规性检查。通过表示学习来提取更具表征能力的特征向量,以提高下游合规性检查任务的性能和效率。

1 相关工作

业务流程合规性检查是一项重要的研究课题,近年来,学者们对此进行了大量研究,业务流程合规性检查方法大致可以分为 3 类,即基于过程模型的合规性检查方法、基于机器学习的合规性检查方法和基于深度学习的合规性检查方法。

1.1 基于过程模型的合规性检查方法

在过程挖掘领域,通常将日志记录中的轨迹与挖掘算法提取的过程模型进行对齐。文献[12]利用过程模型的结构和行为特征,提出一种有效的方法,在寻求最佳对齐的同时减少了整体搜索空间。文献[13]以迭代的方式,使用分解方法平衡对齐质量和计算时间。文献[14]引入一种有效的方法来加速对齐的计算,但是此方法需借助分布式计算技术提高对齐效率。文献[15]提出一种跨组织业务流程模型挖掘方法,其有效、准确地实现了流程模型的挖掘。但是,上述所有基于过程模型的合规性检查方法需要挖掘高质量过程模型,不仅较难实现,而且在进行对齐比较时需要消耗大量时间,所以以上方法存在很大的局限性。

1.2 基于机器学习的合规性检查方法

随着机器学习的发展,基于机器学习的业务流程合规性检查方法得到了广泛的研究。文献[16]提出了一种基于贝叶斯模型的时间合规性检查方法,但该方法只对运行过程中的活动执行时间进行检测。文献[17]通过现有的技术对贝叶斯模型进行拓展,该方法不需要任何关于数据和属性的先验知识,并且可以指出异常的根本原因。文献[18]使用 Word2Vec 进行特征表示,提出一种自主的、进化的在线聚类算法,以实现合规性检查。文献[19]比较了决策树、K 近邻、神经网络在合规性检查中的准确率,结果显示,神经网络具有较优的结果。因此,相比深度学习方法,机器学习在准确率方面还有优化空间,同时机器学习方法需要人工设计和选择特征来输入到模型中,易受主观选择的影响。

1.3 基于深度学习的合规性检查方法

随着深度学习的不断发展,深度学习方法被应用在业务流程合规性检查中并取得了较好效果。文献[20]提出一种基于深度学习编码器的合规性检查方法,该方法与机器学习方法相比提高了合规性检查的能力。文献[21]使用长短时记忆(LSTM)网络对异常进行检测,并在数据处理、网络架构、异常评分等方面进行改进,提高了检测质量。文献[22]采用 LSTM 循环神经网络结合自注意力机制来处理业务流程事件之间的长期依赖问题,并使用变分自编码机(VAE)改进学习模型。

目前,Transformer 在自然语言处理(NLP)领域有着出色的表现。由于 NLP 中语句和业务流程中的轨迹都是序列数据,所以大量学者使用 Transformer 处理业务流程管理中的不同任务。文献[23]提出一种基于注意力机制的业务过程合规性检查方法,该方法通过预测下一事件及属性的概率分布来计算该事件各属性的异常评分,从而实现业务流程的合规性检查,但是该方法需要对不同日志提出不同的评分阈值。文献[24]提出了一种半监督

的分类模型，该模型使用注意力机制进行序列处理，能够从编码器的隐藏状态中获得某些信息，并结合双向 LSTM 实现业务流程合规性检查。

除了在合规性检查方面取得优异成果外，Transformer 在其他任务上也效果显著。文献[25]提出一种业务流程剩余时间预测方法，该方法使用双向循环网络建模，并引入注意力机制自动学习不同事件的权重。文献[26]提出了一种基于 XLNet 的业务流程下一活动预测方法，该方法利用 XLNet 作为预训练模型实现长程记忆。

综上所述，基于 Transformer Encoder 构建的 BERT 模型具有捕捉长距离依赖关系的优势。因此，本文提出一种基于改进 BERT 和轻量化卷积神经网络(CNN)的业务流程合规性检查方法。改进 BERT 模型在计算自注意力机制权重时，融合相对上下文关系，能够更准确地捕捉到序列中不同位置之间的关系，该模块称作融合相对上下文关系的 BERT(CtxtBERT)，可以增强特征向量的表示能力。此外，本文还提出一种基于表示学习的业务流程合规性检查框架，该框架包括数据预处理、特征向量表示、合规性检查 3 个模块。该框架通过学习轨迹的高维表示，利用分类算法进行业务流程合规性检查。

2 基础知识

本文重点使用 BERT 模型与 CNN 模型实现业务流程在线合规性检查。因此，本章简要介绍相关的事件日志概念和深度学习模型。

2.1 事件日志

在业务流程管理中，事件日志是记录组织或系统中发生的事件序列的日志文件，由业务过程的执行轨迹组成。为了方便理解，下面给出事件日志中相关概念的形式化定义。

定义 1(事件) 事件是业务流程的主体，每个事件都具有事件 ID、时间戳等属性，由 $e=(a_{\text{attr},1}, a_{\text{attr},2},\cdots,a_{\text{attr},m})$ 表示。其中，$a_{\text{attr},i}$ 表示事件所具有的属性，$1\leqslant i\leqslant m$ 。

定义 2(轨迹) 由一系列事件组成的序列被称作轨迹，由 $\sigma=\{e_1, e_2,\cdots, e_{|n|}\}$ 表示。其中，e_1，$e_2,\cdots,e_{|n|}$ 为流程实例中的事件，$|n|$ 表示轨迹的长度。

定义 3(事件日志) 事件日志是全部轨迹构成的集合，可以用 $L=\{\sigma_1, \sigma_2,\cdots, \sigma_{|L|}\}$ 表示。

定义 4(轨迹前缀) 轨迹前缀是从轨迹开头开始的前 l 个事件构成的轨迹子序列，可由 $\sigma^l=\{e_1, e_2,\cdots, e_l\}$ 表示。其中，$1\leqslant l\leqslant|n|$ 。

BPIC_2017 事件日志的部分样例如表 1 所示，以该表为例，详细描述上述定义。

表 1 BPIC_2017 事件日志的部分样例

Table 1 Some examples of BPIC_2017 event logs

实例 ID	活动	时间戳
Offer_247135719	O_Create Offer	2016/1/2 17:17
Offer_247135719	O_Created	2016/1/2 17:17
Offer_247135719	O_Sent(online only)	2016/1/2 17:19
Offer_247135719	O_Cancelled	2016/1/2 17:21
Offer_941964966	O_Create Offer	2016/1/2 17:21
…	…	…

每个案例中包括多个事件，将事件日志中的事件根据实例 ID 串联成序列，例如编号为 Offer_247135719 的实例串联事件可以得到轨迹 $\sigma=<$O_Create Offer, O_Created, O_Sent(online only), O_Cancelled$>$，其轨迹前缀有 $\sigma^1=\{$O_Create Offer$\}$、$\sigma^2=\{$O_Create Offer, O_Created$\}$，以此类推。事件日志是由各案例对应轨迹构成的集合。

2.2 BERT

BERT[27] 由 Google 在 2018 年提出，其凭借在 11 种不同自然语言处理任务中的优异表现成为当今最具影响力的语言表示模型之一。BERT 是基于双向 Transformer 架构的训练模型，可以同时考虑左、右两侧的上下文信息。图 1 是 BERT 网络模型结构(彩色效果见《计算机工程》官网 HTML 版，下同)。

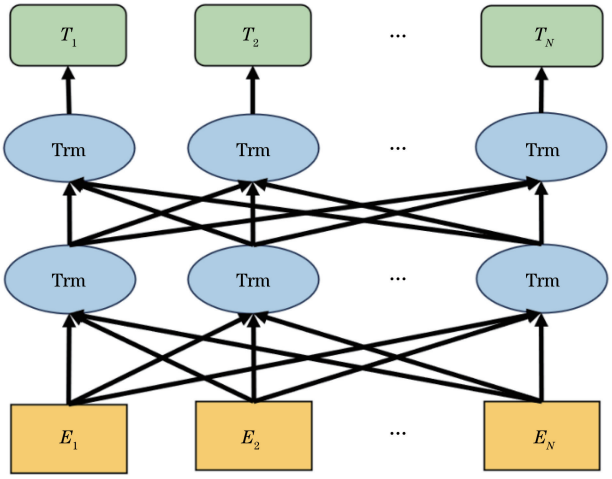

图 1 BERT 网络模型结构

Fig. 1 BERT network model structure

BERT 作为一种强大的预训练语言模型，通过预训练和微调的方式提供高质量向量表示[28]。BERT 模型的预训练包括 2 个任务，即掩码语言建模(MLM)和下一句预测(NSP)。MLM 任务随机覆盖一些输入单词，然后让模型根据上下文来预测

这些被掩盖的单词。在输入序列中给定 2 个句子，NSP 任务预测这 2 个句子是否连续，这个任务有助于学习句子之间的关联性。BERT 模型的输入主要包含标记嵌入、位置嵌入和段嵌入 3 个部分。标记嵌入将各个词转换成固定维度的向量，是一种将词语转成向量的常用方法；位置嵌入用于表示输入序列中标记的位置关系，是一种固定的向量表示；段嵌入也是一组固定的向量表示，每个向量对应一个段落，用于将输入文本中不同段落的标记进行区分，以便将下游任务中存在的两句话同时输入到模型中，例如文本对分类、问答系统等。

2.3 CNN

CNN 是一种深层前馈神经网络，主要由输入层、卷积层、池化层和全连接层等组件构成[29]，通过多层堆叠学习输入数据的特征表示。CNN 最初因为能够有效地从图像中提取出具有层次结构的特征表示而广泛应用于图像处理领域，同时也被引入到自然语言处理领域。CNN 是一种强大的深度学习模型，适用于处理具有网格结构数据的任务，它通过卷积、池化和全连接等操作，可以从输入数据中提取出高层次的特征表示，为图像处理、自然语言处理等领域的任务提供了有效的解决方案。

在图像处理领域，图像是二维数据，图像中的卷积核一般都是正方形的，采用二维卷积，通过从左到右、从上到下滑动窗口进行特征提取。应用在文本序列中，卷积核的宽与特征向量的宽相同，并且卷积核只从上到下滑动窗口来进行特征提取。

3 改进的业务流程合规性检查方法

本文方法的业务流程合规性检查整体框架如图 2 所示，其主要处理流程为：首先，对事件日志进行数据预处理，针对离线事件日志的合规性检查，在历史事件日志的基础上添加异常活动并分别标记所有轨迹的合规性情况，进而构造数据集，针对在线轨迹的合规性检查，在处理历史事件日志时，应当在添加异常活动之前提取出轨迹前缀，进而构造数据集；然后，将轨迹输入表示学习的模型中，输出特征向量；最后，将模型输出的特征向量输入分类预测网络模型中，实现对轨迹的合规性检查。

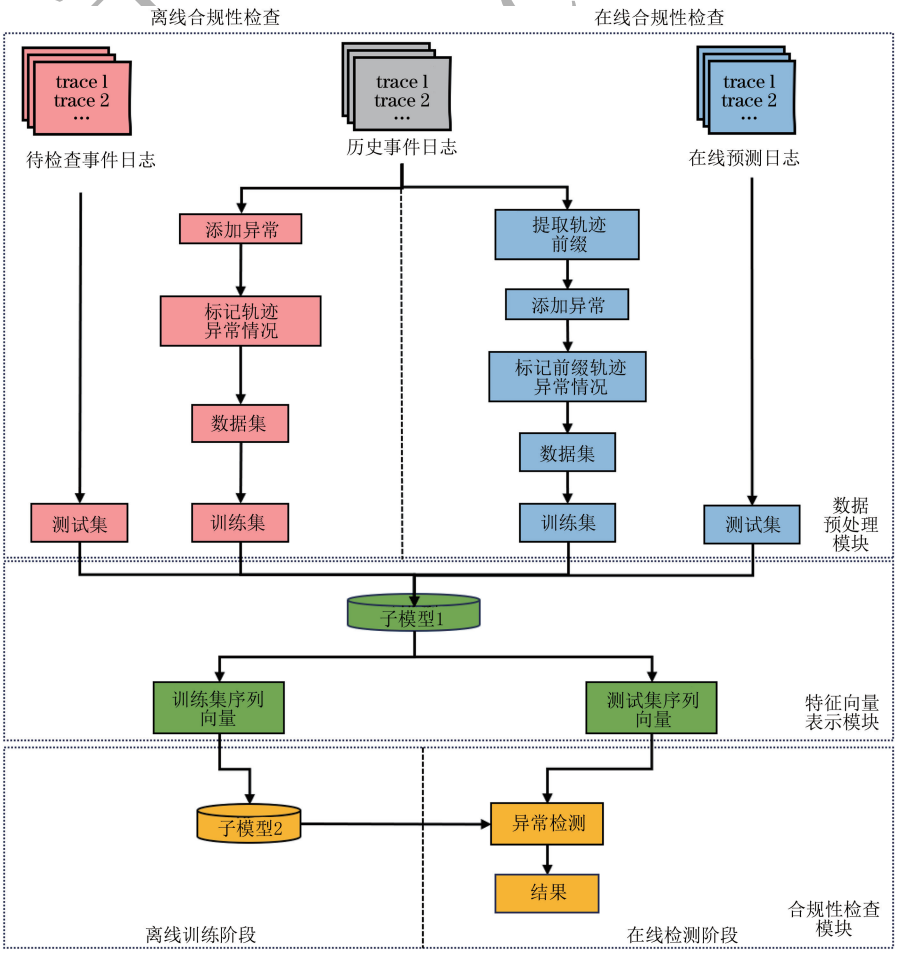

图 2 基于表示学习的业务流程合规性检查框架

Fig. 2 The framework of business process conformance checking based on representation learning

整个框架旨在通过表示学习提取更具表征能力的特征向量，以提高下游预测任务的性能和效果，其优势在于可以自动学习数据中的有用特征，摆脱标签及不同下游任务的束缚，提供更好的预测性能和泛化能力。在该框架的基础上，基于改进 BERT 和轻量化 CNN 的业务流程合规性检查方法，使用 CtxtBERT 模型对预处理的数据集完成特征向量的表示，使用轻量化 CNN 模型实现业务流程合规性检查。该方法可以充分利用 CtxtBERT 模型强大的语义表示能力和轻量化 CNN 对局部特征的敏感性，从而在业务流程合规性检查中获得更好的效果。基于改进 BERT 和轻量化 CNN 的业务流程合规性检查方法结构如图 3 所示。

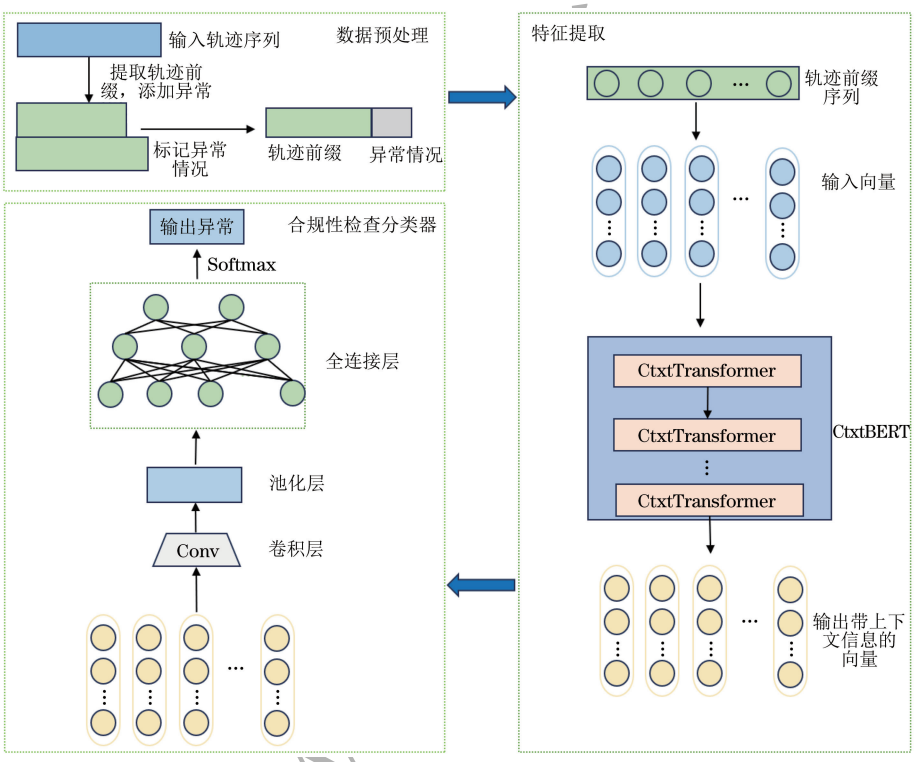

图 3 基于改进 BERT 和轻量化 CNN 的业务流程合规性检查方法结构

Fig. 3 The structure of business process conformance checking method based on improved BERT and lightweight CNN

1）数据预处理。对收集的数据提取轨迹前缀，提取的轨迹前缀标注为正常轨迹，标签标记为 0。使用人工方法将异常插入到轨迹前缀中，随机删除轨迹前缀的活动，与轨迹前缀数据进行对比，若数据集中不存在，则插入数据集中，标注为不拟合轨迹，标签标记为 1。

2）特征向量表示。将预处理后的数据集输入 BERT 模型中，为了捕获相对位置信息，在计算注意力权重时，修改注意力的计算公式，将绝对位置编码与相对位置编码相结合，计算注意力权重。最终，将轨迹序列转化为包含上下文语义信息的高维向量表示。

3）构建合规性检查分类器。将 BERT 模型提取的特征表示作为输入，然后依次经过多层卷积、池化，最后借助全连接层实现合规性检查分类。在卷积层使用不同大小的卷积核获取不同的局部特征，在池化层使用最大池化降低特征维度，在全连接层使用 Softmax 函数实现合规性检查结果的预测。通过设计包含较少层数和参数的 CNN 模型，移除网络中不重要的连接或参数，减小模型大小。

4）模型训练与评估。使用带拟合情况标签的训练集数据对模型进行训练和优化，使用测试集数据对训练模型进行评估。

3.1 数据预处理

以表 1 的事件日志为例，阐述对于事件日志的数据预处理方法：

1）将事件日志中的事件根据实例 ID 串联成序列，得到一条条轨迹序列。

2）轨迹前缀提取是业务流程中实现在线合规性检查的重要步骤。将得到的轨迹按照不同长度划分出轨迹前缀，得到不同长度的前缀序列。

3）在数据集中人工添加异常。具体地，在生成的轨迹前缀序列上随机添加异常活动或删除活动，将该生成序列与事件日志提取的轨迹及轨迹前缀序列进行对比，若不存在，则添加到数据集中。

4)对数据集中的每条轨迹及轨迹前缀标记拟合情况,不拟合轨迹标记为 1,否则标记为 0。

经过上述步骤,得到 BPIC_2017 事件日志的预处理轨迹示例,如表 2 所示。

表 2 BPIC_2017 事件日志的预处理轨迹示例

Table 2 Preprocessing trace example of BPIC_2017 event log

实例 ID	轨迹前缀	合规性情况
Offer_247135719	{O_Create Offer}	0
Offer_247135719	{O_Create Offer, O_Created}	0
Offer_247135719	{O_Create Offer, O_Created, O_Sent(online only)}	0
Offer_247135719	{O_Create Offer, O_Created, O_Sent(online only), O_Cancelled}	0
Offer_247135719	{O_Create Offer, other}	1
Offer_247135719	{O_Create Offer, O_Created, other}	1
Offer_247135719	{O_Create Offer, O_Sent(online only)}	1

3.2 基于 CtxtBERT 的特征向量表示

特征向量表示使用 CtxtBERT 模型进行训练。相对于其他特征向量表示方法,CtxtBERT 模型在多方面存在优势。与 One-Hot 编码方法相比,CtxtBERT 模型可以更好地捕获上下文信息,One-Hot 编码方法对于轨迹中的活动区分度不高;与 Word2Vec 相比,CtxtBERT 模型不仅可以输出词级别的向量表示,而且可以直接输出轨迹序列的向量表示,而 Word2Vec 只能实现轨迹在词级别方面的表示,对于轨迹整体的向量表示,需要通过平均加权或者拼接的方式实现。另外,CtxtBERT 采用了双向建模,不仅考虑了词语左侧的上下文信息,还考虑了右侧的上下文信息,同时添加了轨迹活动之间的相对上下文信息。

对于合规性检查来说,CtxtBERT 主要学习轨迹深层次的特征向量表示,其每次只输入一个轨迹序列。因此,本文主要侧重于使用 MLM 任务来训练 CtxtBERT 模型,帮助模型学习活动之间的上下文信息。在 MLM 任务中,模型的输入无须构建段嵌入,用 $A_{\mathrm{Act},i}$ 表示活动词元,E 是标记嵌入函数,P 是位置嵌入函数,该任务中的输入向量由两部分构成:

1)标记嵌入。对输入轨迹序列每个活动进行嵌入表示,将其转换为向量形式,该向量表示活动的语义信息。

2)位置嵌入。位置嵌入用于表示输入序列的顺序,为每个标记的位置分配一个向量表示。

只需要将每个部分的结果相加即可得到最终的输入向量,将其作为模型的输入。输入向量如式(1)所示:

$$\mathrm{Input\_Embedding}(A_{\mathrm{Act},i}) = E(A_{\mathrm{Act},i}) + P(A_{\mathrm{Act},i}) \tag{1}$$

为了更好地利用上下文信息,对 BERT 模型中的 Transformer 自注意力机制进行改进,在计算注意力时,将绝对位置编码与相对位置编码相结合。传统 Transformer 的自注意力机制基于绝对位置编码进行不同活动之间的位置表示,在计算注意力权重时,这导致每个位置的注意力权重只取决于其与其他活动位置的相关性,而无法区分不同位置之间的相对距离。对查询向量和键向量进行点积计算,然后对其进行缩放,除以键向量的维度,再进行 Softmax 操作,从而获得注意力权重,最后,将权重与值向量相乘得到注意力表示,其计算公式如式(2)所示:

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

式中:$\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 分别代表查询向量、键向量和值向量;d_k 表示键向量的维度。为了更好地编码活动之间的位置信息,添加相对位置编码。添加相对位置编码后,将键向量与相对位置矩阵相加,再与查询向量进行点积计算,随后进行缩放和 Softmax 操作,最后将权重与值向量相乘得到注意力表示,其计算公式如式(3)所示:

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V},\boldsymbol{R}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}(\boldsymbol{K}+\boldsymbol{R})^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{3}$$

式中:$\boldsymbol{R}$ 代表相对位置矩阵。

对比式(2)和式(3),能够看到在计算注意力权重时还会添加相对位置矩阵,使注意力权重能够根据相对位置进行调整。添加相对位置编码后的自注意力机制结构如图 4 所示。

3.3 基于轻量化 CNN 的合规性检查分类器

合规性检查分类器使用轻量化 CNN 模型。本文中使用 CtxtBERT 完成特征向量的表示,得到的句向量包含了丰富的语义信息,可以直接用于业务流程合规性检查分类。因此,本文在构建 CNN 模型时,将生成的特征向量作为输入向量送入轻量化

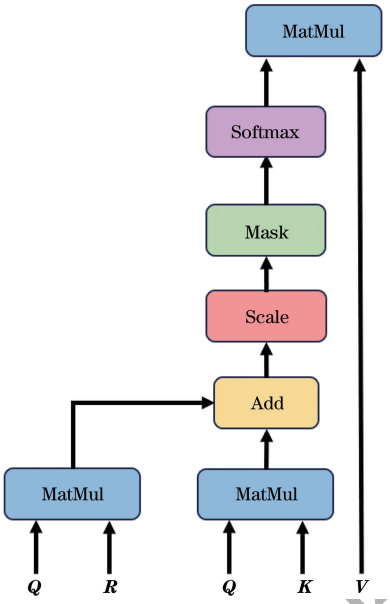

图 4 添加相对位置编码后的自注意力机制结构

Fig.4 The structure of self-attention mechanism after adding relative position encoding

CNN 模型中。

轻量化 CNN 的设计思路主要体现在 3 个方面:1)卷积层的输入通道数为 1,大大降低参数量;2)使用多尺度卷积操作提高网络对输入数据的建模能力,选取较小卷积核,使用大小为 $2\times l$、$3\times l$、$4\times l$(l 代表轨迹向量的维度)的卷积核提取局部特征;3)降低卷积核的数量,使用 128 个卷积核。

减小卷积核的数量和大小,可以降低模型的参数量和计算量,减少模型运行时间。但值得注意的是,根据不同数据集的特点,可以对上述网络参数进行灵活配置与优化,以达到更短的运行时间以及更高的准确率,详细过程如下:

1)模型通过不同大小卷积核捕获不同尺度的特征,目的是在提取特征时发挥各自的优势,提高网络对不同尺度特征的表达能力,有助于检测和理解轨迹中的细微差异和复杂关系,能够从复杂冗长的输入轨迹数据中提取关键信息,更准确地判断轨迹的不拟合情况,卷积核的宽度与轨迹向量的维度相匹配。

2)通过最大池化减少特征向量的尺寸,用于对特征图进行降采样,降低模型复杂度,并且获取特征的最显著信息。

3)在输入全连接层之前进行 Dropout 操作,以一定概率(Dropout 率)随机地将神经元的输出置为零,丢弃无效特征数据,减少数据过拟合现象。

4)通过全连接层将高级特征映射到最终的输出类别。

精心设计的 CNN 网络结构包含较少的参数与层数,能够有效提升模型运算速度。轻量化 CNN 模型结构如图 5 所示。

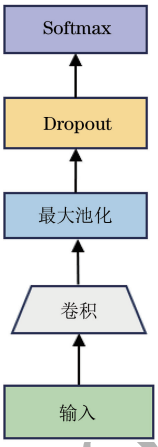

图 5 轻量化 CNN 模型结构

Fig.5 The structure of lightweight CNN model

在卷积层使用大小为 $2\times l$、$3\times l$、$4\times l$ 的卷积核提取局部特征,卷积核的数量为 128 个。卷积操作的计算公式如式(4)所示:

$$c_i = f\left(\sum_{j=i}^{i+h-1} x_i \cdot w_j + b\right) \tag{4}$$

式中:c_i 表示卷积后得到的特征;f 表示激活函数;x_i 是第 i 个词的特征向量表示;w_j 表示卷积核的权重;b 是偏置项;h 代表卷积核的大小。池化操作的计算公式如式(5)所示:

$$p = \max(c_1, c_2, \cdots, c_{n-h+1}) \tag{5}$$

式中:max 表示最大池化操作;p 表示最大池化后得到的最大值;n 表示特征向量的长度;h 代表卷积核的大小。

在全连接层使用 Softmax 激活函数来获得最终的轨迹合规性检查结果。Softmax 激活函数的定义如式(6)所示:

$$p_i = \mathrm{Softmax}(z_i) = \frac{e^z i}{\sum_{k=1}^{K} e^z k} \tag{6}$$

式中:z 为输出向量,其维度为 k。通过 Softmax 函数得到序列所属类别的最大值,即预测合规性类别。合规性检查分类器超参数配置如表 3 所示。

表 3 合规性检查分类器模型的超参数配置

Table 3 Hyperparameter configuration of conformance checking classifier model

参数	配置
卷积核大小	$2\times l$、$3\times l$、$4\times l$
不同大小卷积核个数	128
池化策略	1-maxpool
Dropout 率	0.1
激活函数	Softmax

4 实验分析

为验证本文提出的基于改进 BERT 和轻量化 CNN 的业务流程合规性检查方法的性能，在 5 个公开数据集上进行实验，以准确率(Accuracy)作为评价指标对消融实验和对比实验进行分析。

4.1 实验数据

为了评估不同方法进行在线合规性检查的性能，在 5 个公开数据集上进行分析与评估。Help Desk (https://data.4tu.nl)数据集描述意大利某公司的票务业务流程的相关信息；BPIC_2017 描述荷兰金融机构的贷款申请流程；BPIC_2020 描述 2 年差旅费索赔相关事件；Sepsis Cases 描述医院脓毒症病例事件；Hospital Billing 描述与医院提供的医疗服务计费相关的事件。表 4 为事件日志的相关数据分析。

表 4 事件日志的相关数据分析

Table 4 Related data analysis of event logs

数据集	轨迹数	去重轨迹数	活动数	事件数	最大轨迹长度
Help Desk	4 579	225	14	21 334	15
BPIC_2017	42 995	16	8	193 841	5
BPIC_2020	10 361	97	17	56 420	24
Sepsis Cases	1 050	846	16	15 198	185
Hospital Billing	69 252	1 013	18	451 341	217

4.2 实验设置

实验中的设备处理器为 Intel® Core™ i5-12500H 3.10 GHz，GPU 为 GeForce RTX3060。另外，Python 版本为 3.8，PyTorch 版本为 1.12.1，CUDA 版本为 11.6.0。

实验中模型的参数设置如表 5 所示。使用 BERT 进行特征向量训练时，Epoch 数设置为 30，学习率为0.001，Batch size为16，模型的输出向量维度为最大轨迹长度；在合规性检查分类器进行分类时，Epoch 数设置为 20，Batch size 为 32，学习率为 0.001，Dropout 为 0.1，卷积核个数为 128。

表 5 模型参数配置

Table 5 Model parameters configuration

模型	参数	参数值
BERT 特征向量表示模型	Epoch	30
	Batch size	16
	学习率	0.001
	多头注意力机制	8
合规性检查分类器	Epoch	20
	Batch size	32
	学习率	0.001

4.3 评价指标

本实验以准确率(Accuracy，A)为评估指标比较合规性检查的性能。准确率表示分类模型正确分类样本的比例。在合规性检查分类时，假如数据集中包括 m 个不拟合轨迹和 n 个拟合轨迹，模型对 m 个不拟合轨迹正确预测 x 个，对 n 个拟合轨迹正确预测 y 个，则准确率的计算公式如式(7)所示：

$$A=\frac{x+y}{m+n}\times 100\% \tag{7}$$

4.4 实验结果分析

4.4.1 消融实验

为对比 CtxtBERT 与传统 BERT 模型在特征向量表示方面的优劣，本节设计了消融实验。通过修改模型中自注意力机制的计算方法，在计算时添加相对位置矩阵，评估该部分对模型的优化效果。在传统 BERT 模型的基础上，融合相对上下文关系，保证超参数及模型其他参数相同，下游预测任务使用相同参数的合规性检查分类器，同时使用相同的训练集与测试集，2 种算法的准确率对比如表 6 所示(最优结果加粗标注)。

表 6 2 种算法的准确率比较

Table 6 Comparison of accuracy between two algorithms %

算法	Help Desk	BPIC_2017	BPIC_2020	Sepsis Cases	Hospital Billing
BERT+CNN	78.71	73.98	81.96	89.16	**73.72**
CtxtBERT+CNN	**81.32**	**81.37**	**83.83**	**90.21**	72.13

对比 2 种算法的准确率可知，除 Hospital Billing 数据集以外，CtxtBERT+CNN 模型方法都比传统 BERT + CNN 准确率高。对 Hospital Billing 数据集进行分析，发现虽然本数据集的轨迹最大长度是所有数据集中最大的，但数据集中轨迹长度分布极不均匀，其中，长度为 1 的轨迹占比很大，占数据总数的 1/4。而 CtxtBERT 模型能更准确地捕捉序列中的上下文信息，对相对较长的序列具有更理想的预测效果，相对位置信息可以帮助模型更有效地处理长距离的依赖关系，使模型在处理长序列或复杂序列时具有更优的性能。

4.4.2 对比实验

由于基于深度学习的方法性能优于其他方法，因此本文的对比实验主要进行所提方法与其他深度学习方法的比较。使用上述 5 个事件日志数据集，将本文方法与通用的分类模型 Word2Vec + CNN[30]、Transformer[31]、BERT[32] 进行比较，在各个数据集上准确率结果如表 7 所示。

表 7 不同模型在 5 个数据集上的准确率对比

Table 7 Comparison of accuracy of different models on five datasets %

模型	Help Desk	BPIC_2017	BPIC_2020	Sepsis Cases	Hospital Billing
Word2Vec+CNN	80.96	81.24	77.38	70.29	73.61
Transformer	76.85	72.03	80.16	83.61	68.93
BERT	79.11	76.24	83.60	82.94	**75.45**
CtxtBERT+CNN	**81.32**	**81.37**	**83.83**	**90.21**	72.13

从表 7 可以看出，本文所提方法准确率基本高于其他方法，尤其在 Sepsis Cases 数据集中，本方法获得了非常高的准确率，有更明显的优势。观察并分析 Sepsis Cases 数据集，可以发现事件日志中轨迹序列多为长序列且较为复杂，CtxtBERT+CNN 模型本身擅长处理相对较长的序列，所以在该数据集上取得了最好效果。

4.4.3 轻量化 CNN 效率优势分析

为验证卷积核数量对运行时间的影响，在保证其他参数一致的情况下，使用不同数量的卷积核进行实验，以评估轻量化 CNN 的运行效率。使用相同的硬件环境和训练流程，配置不同数量的卷积核进行实验，并记录模型在训练阶段的性能和运行时间。在各数据集上，不同数量卷积核运行时间对比结果如图 6 所示，不同数量卷积核准确率对比结果如图 7 所示。其中，图 6 比较了 2 种卷积核在不同数据集上不同训练轮次所花费的运行时间。

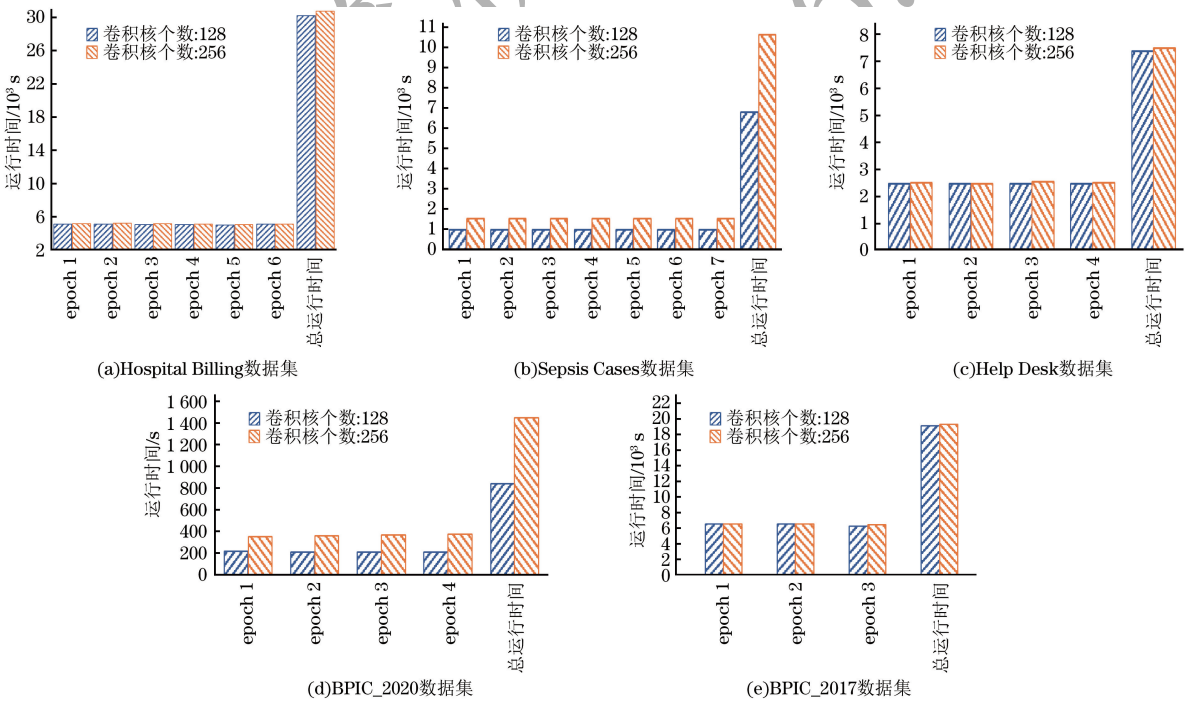

图 6 不同卷积核在 5 个数据集上的运行时间对比

Fig. 6 Comparison of runtime of different convolutional kernels on five datasets

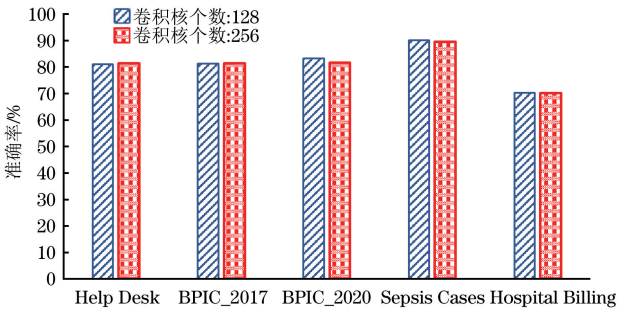

图 7 不同卷积核在 5 个数据集上的准确率对比

Fig. 7 Comparison of accuracy of different convolutional kernels on five datasets

分析实验结果可知，相比使用 128 个卷积核，使用 256 个卷积核需要更长的时间来完成训练，尽管在单位训练轮次下时间增长相对较小，但在整体训练过程中不断累积迭代，导致这种时间差距在完成训练后变得尤为明显，尤其是在包含大量训练样本和复杂特征的数据集上，随着迭代次数的增加，时间差距会更加明显。

5 结束语

本文在基于表示学习的业务流程合规性检查框架的基础上，实现了基于改进 BERT 和轻量化

CNN 的业务流程合规性检查方法。首先，提取事件日志中的轨迹前缀序列，构造带轨迹拟合情况标记的数据集；然后，使用 CtxtBERT 完成特征向量的表示，更好地利用上下文信息；最后，应用轻量化 CNN 模型构建合规性检查分类器，更准确地完成业务流程合规性检查。5 个真实数据集上的实验结果表明本文方法的预测准确率相较对比方法有较大提升。在下一步工作中，将致力于提高模型训练效率并实现多角度的合规性检查。此外，还会研究 CtxtBERT＋CNN 模型在其他业务流程管理任务中的应用。

参考文献

[1] VERA-BAQUERO A, COLOMO-PALACIOS R, MOLLOY O. Business process analytics using a big data approach[J]. IT Professional, 2013, 15(6): 29-35.

[2] MENDLING J, BAESENS B, BERNSTEIN A, et al. Challenges of smart business process management: an introduction to the special issue[J]. Decision Support Systems, 2017, 100: 1-5.

[3] RAMASAMY A, CHOWDHURY S. Big data quality dimensions: a systematic literature review[J]. Journal of Information Systems and Technology Management, 2020, 17: e202017003.

[4] SARKER I H. Machine learning: algorithms, real-world applications and research directions[J]. SN Computer Science, 2021, 2(3): 160.

[5] POPOVIČ A, HACKNEY R, TASSABEHJI R, et al. The impact of big data analytics on firms' high value business performance[J]. Information Systems Frontiers, 2018, 20(2): 209-222.

[6] ARIYALURAN HABEEB R A, NASARUDDIN F, GANI A, et al. Real-time big data processing for anomaly detection: a survey[J]. International Journal of Information Management, 2019, 45: 289-307.

[7] BÖHMER K, RINDERLE-MA S. Multi-perspective anomaly detection in business process execution events[EB/OL]. [2023-09-05]. https://link.springer.com/chapter/10.1007/978-3-319-48472-3_5.

[8] AUGUSTO A, MENDLING J, VIDGOF M, et al. The connection between process complexity of event sequences and models discovered by process mining[J]. Information Sciences, 2022, 598: 196-215.

[9] 沈晓林，刘聪，李会玲，等. 基于流程模型分解的分布式合规性检查方法[J/OL]. 计算机集成制造系统：1-15. [2023-09-05]. https://kns.cnki.net/kcms/detail/11.5946.TP.20230329.1714.003.html.
SHEN X L, LIU C, LI H L, et al. Distributed compliance inspection method based on process model decomposition[J/OL]. Computer Integrated Manufacturing Systems: 1-15. [2023-09-05]. https://kns.cnki.net/kcms/detail/11.5946.TP.20230329.1714.003.html. (in Chinese)

[10] 徐兴荣，张帅鹏，李婷，等. 基于轨迹聚类的业务流程剩余时间预测方法[J]. 计算机工程，2022，48(11)：247-256.
XU X R, ZHANG S P, LI T, et al. Business process remaining time prediction method based on trajectory clustering[J]. Computer Engineering, 2022, 48(11): 247-256. (in Chinese)

[11] KO J, COMUZZI M. Online anomaly detection using statistical leverage for streaming business process events[EB/OL]. [2023-09-05]. https://link.springer.com/chapter/10.1007/978-3-030-72693-5_15.

[12] SONG W, XIA X X, JACOBSEN H A, et al. Efficient alignment between event logs and process models[J]. IEEE Transactions on Services Computing, 2017, 10(1): 136-149.

[13] LEE W L J, VERBEEK H M W, MUNOZ-GAMA J, et al. Recomposing conformance: closing the circle on decomposed alignment-based conformance checking in process mining[J]. Information Sciences, 2018, 466: 55-91.

[14] CHENG L, LIU C, ZENG Q T. Optimal alignments between large event logs and process models over distributed systems: an approach based on petri nets[J]. Information Sciences, 2023, 619: 406-420.

[15] 刘聪，李会玲，曾庆田，等. 跨组织业务流程模型挖掘与质量评估[J]. 计算机学报，2023，46(3)：643-656.
LIU C, LI H L, ZENG Q T, et al. Discovery and evaluation of cross-organization business process models[J]. Chinese Journal of Computers, 2023, 46(3): 643-656. (in Chinese)

[16] ROGGE-SOLTI A, KASNECI G. Temporal anomaly detection in business processes[EB/OL]. [2023-09-05]. https://link.springer.com/chapter/10.1007/978-3-319-10172-9_15.

[17] PAUWELS S, CALDERS T. An anomaly detection technique for business processes based on extended dynamic Bayesian networks[C]// Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. New York, USA: ACM Press, 2019: 494-501.

[18] VERTUAM-NETO R, TAVARES G, CERAVOLO P, et al. On the use of online clustering for anomaly detection in trace streams[C]// Proceedings of the XVII Brazilian Symposium on Information Systems. New York, USA: ACM Press, 2021: 1-8.

[19] NGUYEN H, DUMAS M, LA ROSA M, et al. Business process deviance mining: review and evaluation[EB/OL]. [2023-09-05]. https://arxiv.org/abs/1608.08252v1.

[20] VIJAYAKAMAL M, VASUMATHI D. A novel approach to detect anomalies in business process event logs using deep learning algorithm[EB/OL]. [2023-09-05]. https://link.springer.com/chapter/10.1007/978-981-16-1249-7_34.

[21] LAHANN J, PFEIFFER P, FETTKE P. LSTM-based anomaly detection of process instances: benchmark and tweaks[EB/OL]. [2023-09-05]. https://link.springer.com/chapter/10.1007/978-3-031-27815-0_17.

[22] ELAZIZ E A, FATHALLA R, SHAHEEN M. Deep reinforcement learning for data-efficient weakly supervised business process anomaly detection[J]. Journal of Big Data, 2023, 10(1): 33.

[23] 孙晋永，周博文，闻立杰，等. 基于注意力机制的业务过程异常检测方法[J]. 计算机集成制造系统，2022，28(10)：3039-3051.
SUN J Y, ZHOU B W, WEN L J, et al. Anomaly detection of business processes based on attention mechanism[J]. Computer Integrated Manufacturing Systems, 2022, 28(10): 3039-3051. (in Chinese)

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

[25] 倪维健，孙宇健，刘彤，等. 基于注意力双向循环神经网络的业务流程剩余时间预测方法[J]. 计算机集成制造系统，2020，26(6)：1564-1572.
NI W J, SUN Y J, LIU T, et al. Business process remaining time prediction using bidirectional recurrent neural networks with attention[J]. Computer Integrated Manufacturing Systems, 2020, 26(6): 1564-1572. (in Chinese)

[26] 夏灿铭，邢玛丽，何胜煌. 基于 XLNet 的业务流程下一活动预测方法[J]. 计算机集成制造系统，2023，29(10)：3496-

3503.
XIA C M, XING M L, HE S H. XLNet-based next activity prediction method of business process [J]. Computer Integrated Manufacturing Systems, 2023, 29(10): 3496-3503. (in Chinese)
[27] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2023-09-05]. https://arxiv.org/abs/1810.04805.
[28] WIEDEMANN G, REMUS S, CHAWLA A, et al. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings [EB/OL]. [2023-09-05]. https://arxiv.org/abs/1909.10430v2.
[29] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. [2023-09-05]. https://arxiv.org/abs/1706.03762.
[30] WANG J, TANG Y, HE S, et al. LogEvent2vec: LogEvent-to-vector based anomaly detection for large-scale logs in Internet of Things [J]. Sensors (Basel), 2020, 20(9): E2451.
[31] PIRNAY J, CHAI K. Inpainting transformer for anomaly detection[EB/OL]. [2023-09-05]. https://arxiv.org/abs/2104.13897.
[32] GONZÁLEZ-CARVAJAL S, GARRIDO-MERCHÁN E C. Comparing BERT against traditional machine learning text classification[EB/OL]. [2023-09-05]. https://arxiv.org/abs/2005.13012v2.

编辑 吴云芳