Computer Science and Application 计算机科学与应用, 2022, 12(11), 2573-2583
Published Online November 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1211262

Hans汉斯

基于GAN的高企认定申报数据异常检测模型

陈丽丽[1]，孙　伟[2]，洪英汉[3]，胡　意[1]

[1]广东省科技创新监测研究中心信息规划部，广东 广州
[2]广东工业大学计算机学院，广东 广州
[3]韩山师范学院计算机学院，广东 潮州

收稿日期：2022年10月23日；录用日期：2022年11月21日；发布日期：2022年11月28日

摘　要

高新技术企业认定需要审核的数据繁多，企业在申报过程中经常出现错填、漏填、数据出错、数据不完整等数据异常问题，影响高新技术企业的正常评定程序。通过对广东省高新技术企业认定系统已有数据的研究分析，提出一种基于门控循环神经网络和生成对抗网络的高新技术企业认定申报数据异常检测模型。基于生成对抗网络(GAN)的申报数据异常检测模型在通过生成网络G学习正常样本的分布，使用判别网络D来判别申报数据是不是“真实的”，从而实现数据异常检测。在高新技术企业认定事项管理数据集上进行了实验，实验结果证明了本文提出的模型优于其他模型。

关键词

异常检测，生成对抗网络，门控循环神经网络，高企申报材料

Abnormal Detection Model of High Enterprise Identification and Declaration Data Based on GAN

Lili Chen[1], Wei Sun[2], Yinghan Hong[3], Yi Hu[1]

[1]Department of Information Planning, Guangdong Science and Technology Innovation Monitoring Research Center, Guangzhou Guangdong
[2]School of Computer, Guangdong University of Technology, Guangzhou Guangdong
[2]School of Computer, Hanshan Normal University, Chaozhou Guangdong

Received: Oct. 23rd, 2022; accepted: Nov. 21st, 2022; published: Nov. 28th, 2022

Abstract

There are a lot of data that need to be reviewed for the identification of high-tech enterprises. In

文章引用: 陈丽丽, 孙伟, 洪英汉, 胡意. 基于 GAN 的高企认定申报数据异常检测模型[J]. 计算机科学与应用, 2022, 12(11): 2573-2583. DOI: 10.12677/csa.2022.1211262

the process of declaration, enterprises often have abnormal data problems such as misfiling, missing filling, data error and incomplete data, which affect the normal evaluation procedure of high-tech enterprises. Based on the research and analysis of the existing data of the identification system of high-tech enterprises in Guangdong Province, this paper proposes an anomaly detection model based on the gated recurrent neural network and the generative adversarial network. The reported data anomaly detection model based on Generative Adversarial Network (GAN) learns the distribution of normal samples by generating network G, and uses discriminant network D to determine whether the reported data is "real", so as to realize data anomaly detection. The experiment is carried out on the data set of the identification of high-tech enterprises, and the experimental results prove that the model proposed in this paper is superior to other models.

Keywords

Anomaly Detection, Generating Adversarial Network, Gated Recurrent Neural Network, High Enterprise Application Materials

1. 引言

高新技术企业(简称“高企”)即在国家重点支持的高新技术领域内，持续进行研究开发与技术成果转化，形成企业自主知识产权，并以此为基础开展生产经营活动，在中国境内(不包括港、澳、台地区)注册一年以上的企业[1]。为鼓励高新技术产业发展，促进高新技术企业进行持续的技术研发和科技成果转化，提高企业自主创新能力，2008 年 4 月，科技部、财政部、国家税务总局联合发布了《高新技术企业认定管理办法》。2016 年 1 月，科技部、财政部、国家税务总局联合印发修订后的《高新技术企业认定管理办法》，对高新技术企业认定条件、认定程序、申请条件、税收优惠、监督管理等方面均作了修订，加大了对更多中小微创新型企业的倾斜扶持力度，是供给侧结构性改革的重要举措[2]。

自 2008 年国家颁布《高新技术企业认定管理办法》以来，广东省结合自身科技能力和经济基础，制定了一系列相关配套政策措施[3]。广东省科技厅对高企认定业务资源配置方式、管理流程、结构设置、管理设计等都进行了全面改革[1]。2015 年，广东省科技厅对广东省科技业务综合管理系统进行全面的升级改造，将高新技术企业认定业务全面纳入了升级改造后的广东省科技业务管理阳光政务平台中，按照“横向到边、纵向到底”全覆盖原则和“省市两级平台、三级管理”架构，完善了与国家和地市平台对接，推动了与省直部门数据共享应用[4]，实现了高新技术企业认定、新型研发机构评审、企业研发补助资金核定、科技金融服务、科技项目审批等业务的全流程一体化管理。

高新技术企业认定包括申报通知、企业申报、地市审核、企业提交材料、专家评审认定、复核答辩、认定结果公布等环节[1]，认定程序非常复杂，企业需要填报的数据繁多。由于需要填报的数据繁多，企业填报时容易出现误填、漏填、数据出错、数据不一致等一系列问题。这些问题不仅增加了管理部门工作负担，更严重影响了高新技术企业认定工作的正常进行。

总体来说，数据误填、漏填、出错和不一致等可统一归为数据异常，对误填、漏填、出错和不一致数据的检测统称为异常数据检测。

2. 相关研究

数据异常检测是数据挖掘的一个热门研究方向，其目标是寻找与多数数据对象明显不同的样本点。在数据的分布图中，这些样本点与其他数据点距离较远，所以也被称为离群点[5]。常用的数据异常检测方法一般分为机器学习方法和深度学习方法。传统机器学习算法可以分为以下三类：无监督机器学习、基于聚类算法和基于距离的方法[6]，聚类算法依赖于聚类效果，其主要目的是聚类并不是为了异常检测，对计算开销大。基于距离的方法对于局部异常点不易找到，且较适合高维数据[7]。深度学习方法中主要是利用长短时记忆网络来捕获时序数据之间的依赖关系，从而可以在不同时段对高企申请数据进行异常检测。但一些时间序列异常检测算法缺少定义的异常模式，输入数据的噪声会影响算法的性能[8]。另外，还存在一些基于重构的深度学习算法，Wang 等人[9]引入变分自动编码器，利用编码器的编码和解码结构对异常数据进行深度的检测，但该方法在检测过程会存在一些偏差，不利于数据样本的生成与捕获。基于生成对抗网络的异常检测是一个新兴的研究领域[10]，该模型使用对抗性训练过程建模正常行为，并测量异常评分来检测异常。为了提升高新技术企业申请认定过程中异常数据的检测能力，针对上述方法各自存在的问题，本文提出将门控循环神经网络和生成对抗网络相结合的异常检测模型，通过训练两组网络来构建高企申报材料中企业注册登记表、主要指标情况表和知识产权列表等正常样本模式，检测企业在申报过程中填报的一些异常数据，及时发现并解决问题，提高企认定的质量，完善和提升当前的科技业务审计和监管水平。

3. 异常检测模型

3.1. 申报数据序列异常定义

高企申报数据是具有时序的序列，而时序数据具有数据体量巨大且拥有多变量的时序特性，无法简单通过基于规则的方法来准确判别异常的特点[11]。高企申报数据序列异常是指某个企业在某个申请时间段内的申报数据出现异常点，即申请表数据存在异常模式的子序列。序列异常检测首先构建每个高企认定表中各个认定数据项的正常序列数据，然后根据生成对抗网络构建的模型去检测不符合条件的异常行为序列[12]。首先对 2010~2019 年近 10 年来高企认定登记表中的申报数据进行划分，每个自然年作为一个时间序列，假如 t 时间段内高企申报的数据为 $X=\{X_1,X_2,\cdots,X_t\}$，其中，$X_i=(x_{i1},x_{i2},\cdots,x_{in}),(i=1,2,\cdots,t)$ 为登记表中的每条高企申数据，x_{it} 表示在 i 时刻高企申报数据中的每个字段，其大小为一维向量。序列异常检测的目标是从时序 X 中找到异常序列集合 $A=\{A_1^{seq},A_2^{seq},\cdots,A_m^{seq}\},A_j^{seq}\ (j=1,2,\cdots,m)$ 表示某段时间内登记表中每条高企数据的异常数据项字段序列。基于 GAN 的高企申报数据异常检测的方法是构建高企申报表中正常数据序列模式，通过正常序列模式识别申报数据中存在的异常序列数据。

3.2. 基于 GAN 网络的异常检测模型

3.2.1. 基本原理介绍

对抗生成网络的原理就是通过生成网络 G (Generator)和判别网络 D (Discriminator)的不断博弈，从而使 G 学习到正常数据的分布[13]。生成网络 G (Generator)能够接收一个随机噪声，通过这个噪声生成逼真的高企申报数据样本达到欺骗判别网络的目的；判别网络 D (Discriminator)用来判别一个高企数据是不是“真实的”。在训练阶段中，生成网络 G 生成逼真的高企申报数据去欺骗判别网络 D，D 根据判别函数辨别 G 生成假的申报数据和真实的数据之间的差距。模型的学习过程就是 G 和 D 构成的一个动态“博弈过程”[14]，经过若干次迭代之后最终达到平衡。网络的整体结构如图 1 所示。

本模型采用门控循环单元(GRU)作为基础网络，GRU 是循环神经网络的一种变体。相比长短期记忆网络，使用 GRU 能够更好的捕获时序数据的特征，并且由 GRU 具有结构具有参数量少、易于训练等特

点，因此，本文选用 GRU 作为基础网络来提高模型训练效率。GRU 基础网络结构如图 2 所示。

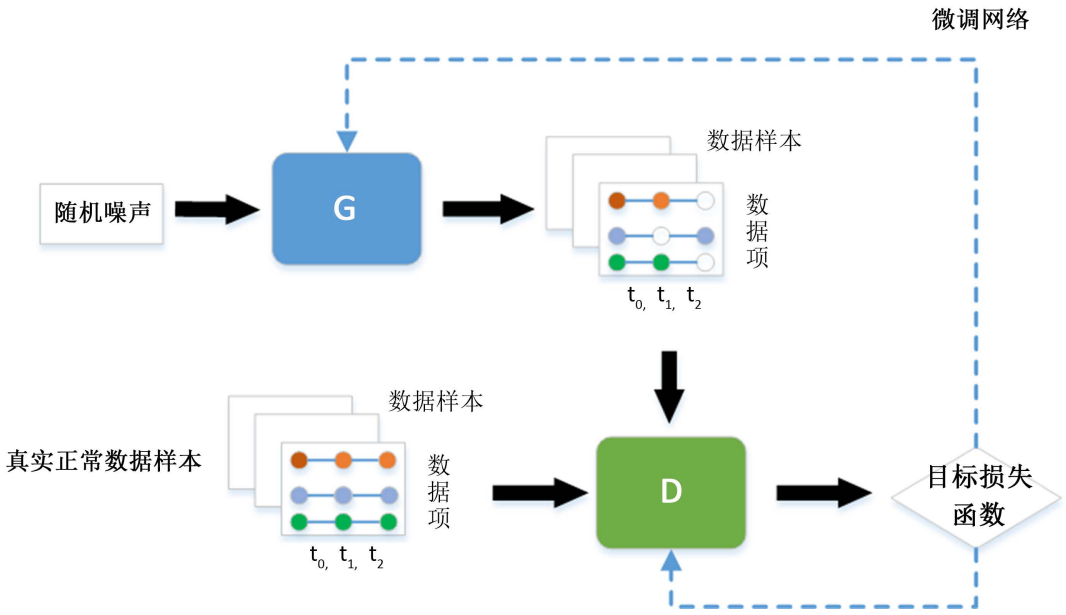

Figure 1. GAN network architecture diagram
图 1. GAN 网络架构图

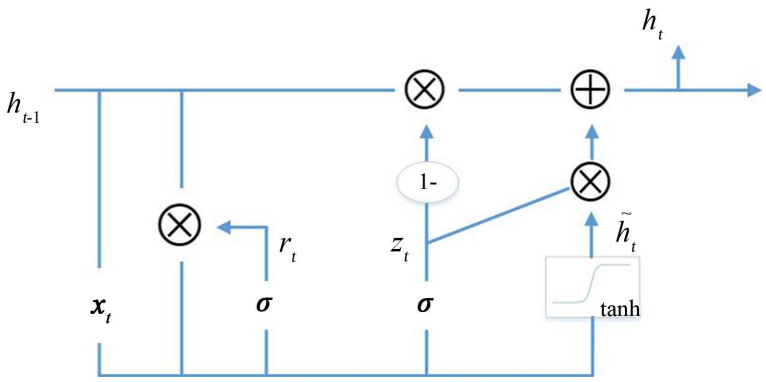

Figure 2. GAN gating diagram
图 2. GRU 门控图

从 GRU 门控网络图中可以看出，在 t 时刻，GRU 隐藏层的输入向量为 x_t，输出向量为 h_t，记忆单元为 $\tilde{h}_t$。其公式可以形式化为(1)所示：其中，

$$\begin{aligned} z_t &= \sigma\left(W_z \cdot \left[h_{t-1}, x_t\right]\right) \\ r_t &= \sigma\left(W_r \cdot \left[h_{t-1}, x_t\right]\right) \\ \tilde{h}_t &= \tanh\left(W \cdot \left[r_t * h_{t-1}, x_t\right]\right) \\ h_t &= \left(1 - z_t\right) * h_{t-1} + z_t * \tilde{h}_t \end{aligned} \tag{1}$$

其中，W 表示权重，σ 是 sigmod 函数，tanh 是正切函数。通过更新门控制前边记忆信息能够继续保留到当前时刻的数据量，即决定有多少前一时间步的信息和当前时间步的信息要被继续传递到未来；重置门则控制要遗忘多少过去的信息。

3.2.2. 异常检测网络目标函数

经典对抗生成网络的损失函数采用 Wasserstein 对抗损失[8]，如式(2)所示：

$$WL = E_{x\sim P_X}\left[\log D_x\left(x\right)\right] + E_{z\sim P_Z}\left[\log\left(1-D_x\left(G_2\left(z\right)\right)\right)\right] \tag{2}$$

其中，E 表示期望，$D_x\left(x\right)$ 为判别映射函数，$G_2\left(z\right)$ 为生成映射函数，P_X 为真实数据分布，P_Z 为原始噪声分布。

由于模型可能存在梯度爆炸的风险，为了避免这个问题，本文采用 Wasserstein 损失函数 WL 来训练网络的目标损失，针对两个生成函数和判别函数分别设计目标损失函数，使得模型能够更好地进行异常检测。

对于生成函数 $G_1: X \to Z$ 和判别函数 $D_z: Z \to P$，目标损失函数 V_Z 设为：

$$\min_{G_1}\max_{D_z\in\hat{D}_z} V_Z\left(D_z, G_1\right) \tag{3}$$

其中，$V_Z\left(D_z, G_1\right) = E_{x\sim P_X}\left[D_z\left(G_1\left(x\right)\right)\right] - E_{z\sim P_Z}\left[D_z\left(z\right)\right]$。

$\hat{D}_z$ 表示 1-Lipschitz 连续函数集，即：对 $\forall x_1, x_2 \in domf$

$$\left\|f\left(x_1\right) - f\left(x_2\right)\right\| \le \left\|x_1 - x_2\right\| \tag{4}$$

本模型对连续函数值的上界进行约束，使得网络权重在更新过程中不会发生激烈变化，避免模型在优化过程中发生梯度爆炸的风险；同时，判别函数将生成的逼真高企申报数据与真实认定数据进行拟合判断，得到模型训练结果。

对于生成函数 $G_2: Z \to X$ 和判别函数 $D_x: X \to P$，目标函数损失 V_X 设为：

$$\min_{G_2}\max_{D_x\in\hat{D}_x} V_X\left(D_x, G_2\right) \tag{5}$$

其中，$V_X\left(D_x, G_2\right) = E_{x\sim P_X}\left[D_x\left(x\right)\right] - E_{z\sim P_Z}\left[D_x\left(G_2\left(z\right)\right)\right]$。

采用生成对抗网络模型有利于时序数据的重构，为了平衡以上两个损失函数的搜索空间，引入循环一致性损失函数进行约束，如式(6)所示：

$$V_{L2}\left(G_1, G_2\right) = E_{x\sim P_X}\left[\left\|x - G_2\left(G_1\left(x\right)\right)\right\|_2\right] \tag{6}$$

其中，$\left\|\bullet\right\|_2$ 表示原始样本与重构样本 $\tilde{x}$ 差的 L2 范式。

综合上式可得到对抗生成网络的目标损失函数如式(7)所示：

$$\min_{\{G_1,G_2\}}\max_{\{D_z\in\hat{D}_z, D_x\in\hat{D}_x\}} V_Z\left(D_z, G_1\right) + V_X\left(D_x, G_2\right) + V_{L2}\left(G_1, G_2\right) \tag{7}$$

3.2.3. 训练过程

首先，从标准正态分布中采样得到随机时序噪声，将噪声输入生成器 G 中，用以生成假的高企申报数据；然后，使用判别网络判别真实样本数据和 G 生成的假数据；最后，通过计算由生成函数和判别函数组成的目标函数对网络进行微调，直至达到纳什平衡。

生成网络 G 由 $G_1: X \to Z$ 和 $G_1: X \to Z$ 两个映射函数构成，函数 G_1 对高企申报真实训练样本 X 降维到子空间 Z 中，函数 G_2 是将子空间 Z 随机生成 X，Z 服从 $N\left(0,1\right)$ 多元正态分布，通过 G_1 和 G_2 两个映射函数对 X 进行重构，即：

$$\forall X_i \in X,\ X_i \to G_1\left(X_i\right) \to G_2\left(G_1\left(X_i\right)\right) = \hat{X}_i\left(1\right) \tag{8}$$

判别网络 D 由 $D_x: X \to P$ 和 $D_z: Z \to P$ 所构成，D_x 用来判断真实数据样本的概率为 P，尽可能判别

真实数据和随机生成数据样本之间的差距；D_z 判定给定的向量 Z 编码得到真实数据样本的概率 P，区分编码生成的潜在向量和随机样本之间的差距。同时，网络模型还将引入 Mask 层和 Dropout 层解决异常数据拟合问题，提高训练网络的稳定性。

3.3. 异常检测模型

在进行多轮的迭代训练后，将训练好的鉴别器和生成器用于申报数据异常检测。异常检测部分以真实高企申报数据测试样本 X_{Test} 为输入，然后分别计算生成网络的重构损失 RL 和判别网络的判别损失 DL，然后根据计算的异常分数来判断是否为异常数据序列。异常检测过程如图 3 所示。

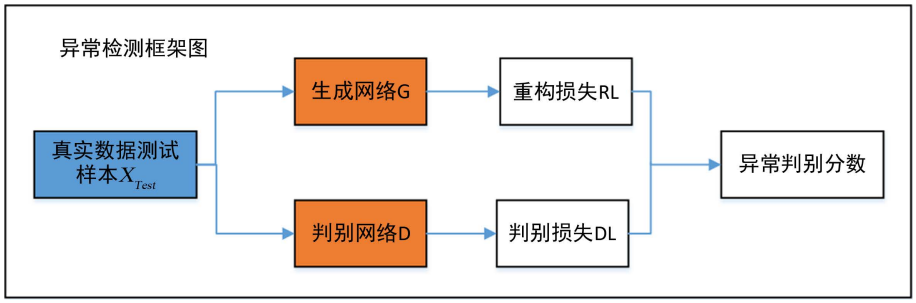

Figure 3. Anomaly detection process
图 3. 异常检测过程

本文提出的高企申报数据异常检测算法描述如下：

输入：高企申报数据样本 $X=\{X_1,X_2,\cdots,X_t\}$
输出：异常数据序列 $A=\{A_1^{seq},A_2^{seq},\cdots,A_m^{seq}\}$

Step 1：对 2010~2019 年来的高企申报数据按自然年进行划分构成一个时间序列，然后使用滑动窗口将训练样本 XTrain 划分为子序列，并设置相应的滑动窗口值为 m，训练样本序列分解为 $X_{1\sim m},X_{m+1\sim 2m},\cdots$；
Step 2：生成正态分布的随机噪声序列 $Z=G(X)$，设置为：$Z_{1\sim m},Z_{m+1\sim 2m},\cdots$；
Step 3：训练 GAN 网络，首先对真实训练样本 X 生成潜在空间，然后对其进行重构；随机噪声 Z 生成伪时序数据，根据目标损失函数 $\min\limits_{\{G_1,G_2\}}\max\limits_{\{D_z\in\hat{D}_z,D_x\in\hat{D}_x\}}V_Z(D_z,G_1)+V_X(D_x,G_2)+V_{L2}(G_1,G_2)$ 进行网络的调整，得到模型所需的参数；
Step 4：数据异常检测，首先将测试样本 XTest 划分为相对应的子序列，然后分别计算重构损失 RL_{Test}，计算判别损失 DL_{Test}，计算异常分数 $Score=\alpha RL_{Test}+(1-\alpha)DL_{Test}$，根据异常分数输出异常数据序列 $A=\{A_1^{seq},A_2^{seq},\cdots,A_m^{seq}\}$。

4. 实验结果与分析

4.1. 实验设计

4.1.1. 实验内容

对于本文所提出的 GRU-GAN 异常检测算法，从给定高企认定事项管理实验数据集上验证性能，与 K-Means、LSTM、Auto-Encoder 3 个异常检测算法进行性能分析。K-Means 算法是一种基于聚类的方法，采用计算数据每个元素到聚集中心点的距离方式判断异常值。LSTM 算法是一种基于深度学习的异常检测方法，它能够有效识别出时间序列中的异常数据，降低计算复杂度。Auto-Encoder 是一种基于自动编

码器的异常检测算法方法，它使用重建误差作为异常评分。

4.1.2. 评价指标

本实验将会在以下四个检测指标出发来衡量模型的可行性及可靠性，准确率(Accuracy, Acc)、精确率(Precision, Pre)、召回率(Recall, Rec)和 F1 分数四个评估指标的计算公式分别如下所示：

$$\mathrm{Acc}=\frac{\mathrm{TN}+\mathrm{TP}}{\mathrm{TN}+\mathrm{TP}+\mathrm{FN}+\mathrm{FP}} \tag{9}$$

$$\mathrm{Pre}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{10}$$

$$\mathrm{Rec}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{11}$$

$$\mathrm{F1}=\frac{2*\mathrm{Pre}*\mathrm{Rec}}{\mathrm{Pre}+\mathrm{Rec}} \tag{12}$$

其中，TP 表示结果为正样本，实际也为正样本，即正样本被正确识别的数量；FP 表示预测结果为正样本，实际为负样本，即误报的负样本数量；TN 表示结果为负样本，实际为负样本，即负样本被正确识别的数量；FN 表示结果为负样本，实际为正样本，即漏报的正样本数量。

准确率一般用来预测算法的检测结果，但有时候会存在数据样本分布不平衡的情况，正常样本类别会影响准确率的生成，本文将通过其他一些指标来评判；精确率可以衡量算法正确检测正样本精度的能力；召回率可以正确预测为正的占全部实际为正的比例；为了平衡精确率和召回率的相互影响，引入 F1 分数来综合评估 GRU-GAN 模型的异常检测性能。

4.1.3. 实验数据

为了验证本研究所构建的基于 GAN 的高企申报数据异常检测算法的科学性和有效性，选取 2010~2019 年广东省高新技术企业认定登记表(表 1)中的申报数据做实证分析，数据来源为广东省科技厅，申报高新技术企业的数量为 1000 家企业样本。实验选取申报登记表中的企业注册登记表(B1)、知识产权详细汇总表(B2)、研发活动情况表(B3)、成果转化表(B4)这四个子表的数据作为异常数据样本检测研究，以评价算法的异常检测性能。对上述表进行数据清洗以及特征预处理等操作，以此得到较为正常的实验数据作为实验正样本；异常数据选取历史检测工作中通过预警记录、人工筛查等方式确认的部分异常数据。最后将正常数据的 80%作为实验训练集，20%正常数据和异常数据共同作为测试集。

Table 1. High enterprise declaration and registration table
表 1. 高企申报登记表

序号	数据子表	字段项
1	企业注册登记表	企业名称、注册时间、企业类型、所属行业、企业规模、行政区域、通信地址、企业是否上市、是否高新区内企业、获得知识产权数量、职工总数、科技人员总数、企业总收入、高技产品收入等
2	知识产权汇总表	知识产权编号、类别、授权日期、获得方式、是否为有效知识产权、相关的核心技术领域等
3	研发活动表	研发活动编号、研发活动、起止时间、技术领域、项目核心技术关键词、技术来源、研发经费总预算等
4	成果转化表	科技成果名称、成果类型、成果来源、成果技术领域、转化形式、转化结果、转化时间等

4.1.4. 实验环境

实验环境配置如下：硬件配置 i7-8700K 处理器，32 G 内存；PyCharm 开发环境，TensorFlow 深度学习框架，使用 Python 编程语言编程。

4.1.5. 实验参数设置

GAN 中的生成网络和判别网络分别采用深度为 20 的 GRU。批量处理数据量设为 50，模型的迭代次数设为 1000 次，网络学习率设为 0.001，模型采用 Adam 进行优化，并且选择 ReLU 激活函数进行网络的训练。不同时间窗口长度大小对算法性能有一定的影响，本文将对原始时序采用滑动窗口进行划分，在实验中将窗口大小设置为 2 的倍数，分 5 组进行交叉实验，窗口大小为 $2*k\left(k=1,2,\cdots,6\right)$，潜在空间的维度设置为 20。

4.2. 结果分析

4.2.1. 异常检测性能对比

分别采用 K-Means、LSTM、Auto-Encoder 和 GRU-GAN 异常检测算法识别 B1、B2、B3、B4 这些子表中的各个字段数据进行异常检测，每个异常检测性能指标如下图所示。

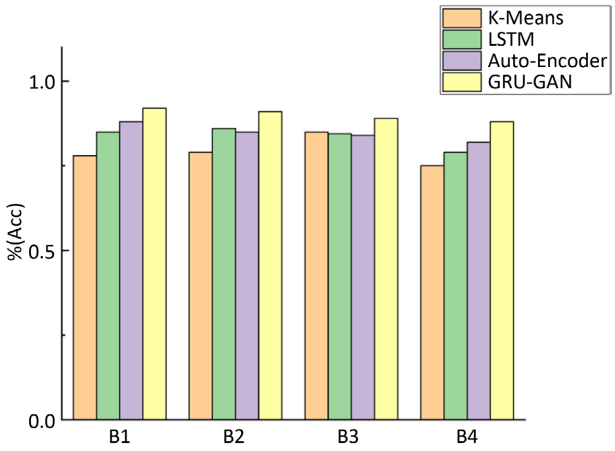

Figure 4. Accuracy comparison chart
图 4. 准确率对比图

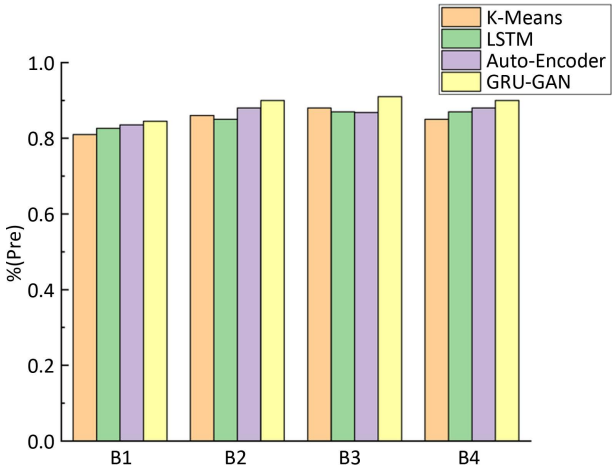

Figure 5. Precision comparison diagram
图 5. 精确率对比图

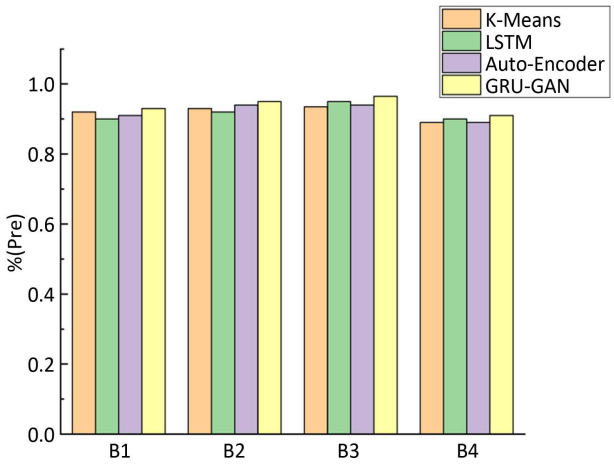

Figure 6. Recall ratio comparison chart
图 6. 召回率对比图

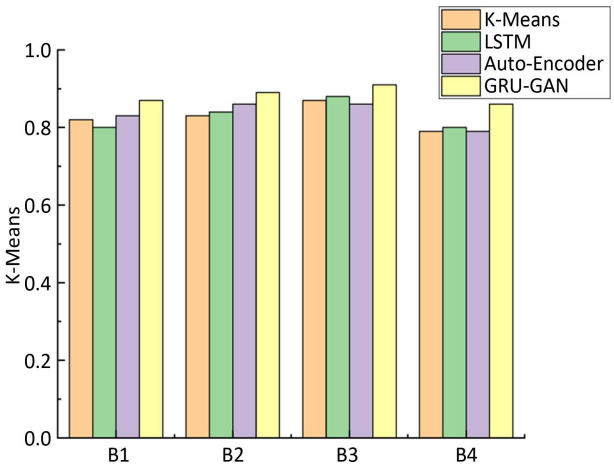

Figure 7. F1 score comparison chart
图 7. F1 得分对比图

实验结果表明，GRU-GAN 的异常检测准确率最高，比 Auto-Encoder 约高 2.6%，比 LSTM 约高 1.9%，比 K-Means 约高 4.2%，由图 4 可以看出，GRU-GAN 在识别 B1 的异常序列中准确率达到 95.8%。对于精确率指标，如图 5 所示，GRU-GAN 的精确率比 K-Means、LSTM 和 Auto-Encoder 方法高三到五个百分点，约为 90%。召回率指标由图 6 得出，GRU-GAN 的召回率高于其他对比方法，并且在识别 B1 的异常序列中召回率达到 94.2%。同样 F1 分数指标也是 GRU-GAN 取得最好的效果。综合以上实验结果可知，由于 GRU-GAN 结合 LSTM 和 Auto-Encoder 两者的优势，GRU-GAN 在四个指标中的得分值都是最高的，比 K-Means、LSTM 和 Auto-Encoder 高出约 2%~6%，促进异常检测性能的提升。

4.2.2. 目标函数分析

图 7 显示了判别损失 V_x、V_z 以及重构损失 V_{L2} 随着网络迭代次数对整个网络的目标函数值变化的影响。

从图 8 可得出，刚开始迭代的时候，目标损失值会迅速下降，第 2 次往后迭代开始，目标函数损失值在趋于平稳的情况下进行下降直到达到收敛。在三个损失函数约束下，异常检测模型展现出稳定的性能，进一步验证了本文所提数据异常检测模型的稳定性。

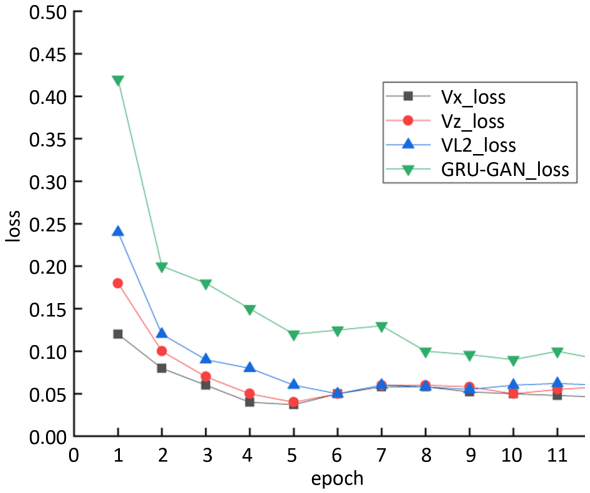

Figure 8. Graph of the change of objective function value with the number of network iterations
图 8. 目标函数值随网络迭代次数变化图

5. 结论

本文为了解决企业申报过程中出现的一些数据异常问题，对广东省高新技术企业认定系统已有数据研究分析，提出一种基于门控循环神经网络和生成对抗网络的高新技术企业认定申报数据异常检测模型。本模型设计生成网络、判别网络以及重构网络来循环构建数据样本，利用生成对抗网络来避免拟合异常数据的风险。实验结果表明，本文提出的异常检测算法有利于异常数据的检测，提高了异常检测的能力，并且有效避免了网络梯度爆炸的问题。本文提出的算法模型较好地解决了高企认定申报材料过程中的一些异常数据检测问题，可以对高企认定填报过程以及认定申请之后带来实时的数据支持。在未来的研究中探索能否结合其他神经网络提高模型效率，并且使训练过程更加稳定的异常检测方法。

基金项目

广东省软科学研究计划项目：基于广东“数字政府”平台建设的科研审计与监管研究——以高新技术企业认定事项管理为例(2020B1010010010)、广东省科技计划(项目)管理风险防范机制研究(2020A1010020007)。

参考文献

[1] 陈辛. 国家高新技术企业认定系统的设计与应用——以广东省为例[D]: [硕士学位论文]. 广州: 广东工业大学, 2017.

[2] 陈之瑶, 罗军, 黄海滨. 促进广东省高新技术企业发展的政策研究[J]. 广东科技, 2016, 25(14): 82-85.

[3] 徐倩. 科技创新驱动发展——广东省高新技术企业培育和认定政策解读[J]. 广东饲料, 2016, 25(7): 15-18.

[4] 王晓湘, 刘洞天, 刘南江, 丁一, 姜立新. 基于 LSTM 的震后通信数据异常检测分析[J]. 中国地震, 2022, 38(2): 270-279.

[5] 解峰, 蔡江辉, 杨海峰, 荀亚玲. 一种基于邻近性和团的异常数据检测算法[J]. 计算机与数字工程, 2021, 49(5): 971-976.

[6] 鲁统伟, 徐子昕, 闵锋. 基于生成对抗网络的知识蒸馏数据增强[J]. 计算机工程, 2022, 48(4): 70-80. https://doi.org/10.19678/j.issn.1000-3428.0060395

[7] 韩来平, 李榕, 张萌. 科研审计与监管: 科学与政治的有机边界活动[J]. 科研管理, 2017, 38(11): 88-94.

[8] 王凤芹, 高龙, 徐廷学, 王丽娜. 基于 LSTM-GAN 的无人机飞行数据异常检测算法[J]. 中国惯性技术学报, 2022, 30(2): 264-271. https://doi.org/10.13695/j.cnki.12-1222/o3.2022.02.019

[9] Habler, E. and Shabtai, A. (2018) Using LSTM Encoder-Decoder Algorithm for Detecting Anomalous ADS-B Messages. *Computers & Security*, **78**, 155-173. https://doi.org/10.1016/j.cose.2018.07.004

[10] 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.

[11] 赵飏, 李晓, 马博, 王保全, 周喜. 基于 LSTM-GAN 的加油时序数据异常检测[J]. 计算机应用与软件, 2022, 39(7): 13-19.

[12] 于冰, 丁友东, 谢志峰, 等. 基于时空生成对抗网络的视频修复[J]. 计算机辅助设计与图形学学报, 2020, 32(5): 769-779.

[13] Zhang, D.Y., Jie, S., Hu, C. and Gao, L.L. (2017) Sharp and Real Image Super-Resolution Using Generative Adversarial Network. In: Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, E.S., Eds., *International Conference on Neural Information Processing*, Vol. 10636, 217-226. https://doi.org/10.1007/978-3-319-70090-8_23

[14] Saxena, D. and Cao, J. (2021) Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions. *ACM Computing Surveys*, **54**, 1-42. https://doi.org/10.1145/3446374