文章编号:1000-5862(2020)04-0385-09

小样本数据生成及其在异常检测中的应用

卢逸君[1,2],滕少华[1*]
(1. 广东工业大学计算机学院,广东 广州 510006;2. 广东省信息安全测评中心,广东 广州 510095)

摘要: 在不平衡数据的应用中,少量的负样本(异常数据)往往是检测准确率低的重要原因,如在主机异常检测领域中,异常样本过少使得检测效果不佳. 为解决这一问题,该文改进了深度卷积生成对抗网络,使其更易于收敛和生成样本. 再通过将改进的深度卷积生成对抗网络用于入侵检测评测数据集 ADFA-LD 异常样本的训练,构造出更多的异常样本. 最后,为验证生成样本的效果,以多种异常检测方法检测对上述增加样本后的平衡数据进行实验,实验结果发现新增加的异常样本能被全部检测出,而且已测出的异常样本无漏检,实现了高检测率和低误报率. 对比实验表明该文提出的小样本数据生成方法能有效解决某些数据不平衡的应用问题.

关键词: 卷积神经网络; 生成式对抗网络; 样本生成; 主机入侵检测; 神经网络
中图分类号: TP 183 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2020.04.10

0 引言

在某些异常检测领域中,由于异常样本较正常样本构造难度更大,从而不论在实验环境下还是在现实环境下,异常样本的比例都往往较低,如主机入侵检测领域的主流数据集(如 ADFA、UMN 所公布的异常样本数量)远远小于正常样本,因此学者在进行研究时经常面临正常和异常样本数量不平衡的问题. 在不平衡的数据下,可能会出现较低检测率和较高误报率,从而影响对分类方法的评价,因此数据不平衡是研究者需要考虑的问题.

对于正常样本数量远远超过异常样本的问题,从数据层面的解决思路主要有 3 类:(i) 降采样,即减少正常样本数量,使之与异常数据大体平衡,这种方法对于需要大量训练数据的检测方法来说,在移除数据的过程中可能造成重要特征的损失,但在处理时间方面有较大优势;(ii) 过采样,即构造少数类样本,传统方法有随机过采样、SMOTE 等,随机过采样会造成过拟合问题,一些样本可能在随机化过程中被重复地模拟; SMOTE 方法能生成噪声数据,减少过拟合,适用于连续型数据,但对于较小的样本类别过采样的效果却不佳[1];(iii) 混合降采样与过采样的方法,即对多数类样本进行降采样,对少数类样本进行过采样,如 SMOTETomek、SMOTEENN,即先用 SMOTE 进行过采样,然后通过 Tomek、ENN 等方法对数据集进行降采样[2].

在异常检测研究中,学者常采用基于 SMOTE 的过采样方法,但该方法适用于连续数据而非离散数据,在一些序列化的数据中并不能直接适用. 因此,本文尝试研究一种基于改进的生成对抗网络的小样本数据生成方法,通过在主机异常检测领域中的样本生成和异常检测实验,验证该算法的有效性.

1 相关工作

作为最经典的过采样方法,SMOTE 于 2002 年由 V. Nitesh 等[3]提出,其策略是对每个少数类样本 a,从它的 N 个最近邻中随机选一个样本 b,然后在 a、b 之间随机选一点作为新合成的少数类样本. Liu Kaijian 等[4]使用 SMOTE 和降采样技巧,通过 PCA 降维、循环神经网络和 k-means 等技术在 KDD99 数据集中实现较好的效果,将异常样本检测的误报率降低了 16%. 但是,SMOTE 过采样方法并不适用于

收稿日期:2020-01-23
基金项目:国家自然科学基金(61702110,61772141,61972102),广东省重点领域研发计划(2020B010166006),广东省教育厅课题(粤教高函[2018]179 号,粤教高函[2018]1 号)和广州市科技计划课题(201903010107)资助项目.
通信作者:滕少华(1962-),男,江西南昌人,教授,博士,主要从事大数据、数据挖掘、数字音频分析与处理、网络安全方面的研究. E-mail: shteng@gdut.edu.cn

主机序列这类异或分布的数据,在主机序列入侵检测领域中,系统调用命令本身并无大小和顺序关系,且命令离散分布,无法通过最近邻关系合成同属一类的少数类样本,因而并不适合通过基于 SMOTE 的方法来扩充样本.

此外,有学者结合多种数据采样技术的优点,构建鲁棒性较强的数据平衡框架. 如 C. Promper 等[5]借鉴了智能电网的思路,针对不平衡的数据问题,为入侵检测系统构造了一个由多种重新采样技术组成的 3 层智能网格通信系统,改善了检测结果.

随着深度学习技术的发展和应用,在近年的研究中,学者开始将基于生成式对抗网络(Generative Adversarial Networks,GAN)的方法纳入异常检测领域数据平衡的研究. J. Lee 等[6]在 CICIDS 2017 数据集中使用 GAN 对少数类样本进行数据平衡,通过随机森林模型的异常检测方法进行检验,发现使用 GAN 的方法能较 SMOTE 大幅提高检测准确率、召回率和 F 值. 在主机序列检测领域中,M. Salem 等[7]采用 CycleGAN 对 ADFA-LD 数据集生成异常样本,将生成的异常数据与真实数据混合,用多层感知机进行训练和分类,通过将该方法生成样本的结果与 SMOTE 过采样的生成结果进行对比,发现该方法能有效降低误报率. 但是,笔者认为该研究不足之处有:(i)采用不恰当的参照对比方法,如前文所述,SMOTE 不适用于 ADFA 主机序列数据集;(ii)生成异常样本的效果评价问题,通过生成异常数据来重新平衡数据集,将召回率从不平衡数据的 17% 提高到 80%,表现并不够理想,甚至略低于 SMOTE 方法;(iii)在样本处理过程中的问题,用 255 来填充序列使之达到统一长度,但由于 ADFA-LD 数据集实际最大序列编号达到 340,序列图像化表示可能失真,且灰度表示方法不够直观;(iv)训练成本高,迭代次数过多,并且笔者判断该实验存在过拟合的可能性较大. 不论如何,该研究反映出利用 GAN 来生成主机序列异常样本具有较大的研究潜力,启发笔者在 GAN 的基础上进一步研究异常样本生成.

2 小样本数据生成及其应用

本文针对小样本数据下的数据不平衡问题,在主机序列入侵检测领域中基于深度卷积生成对抗网络生成异常样本,对训练过程进行优化,以生成更多收敛状态的异常样本,从而改善主机序列数据的不平衡问题,最后通过多种主机异常序列检测方法对生成样本结果进行验证. 完整研究框架分为样本生成与结果验证 2 个部分,以基于特征命令提取的检测方法为例,架构如图 1 所示.

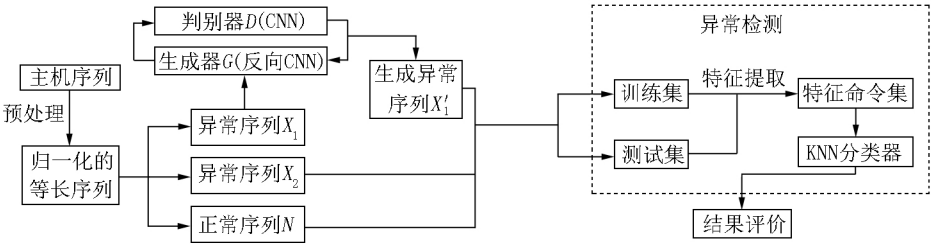


图 1 基于 DCGAN 生成主机异常序列及其检测框架

本文方法的特点有:(i)在输入数据上,对主机序列进行归一化预处理并转换为 $N \times N$ 的 gist_ncar 热力图,能更好地表现序列连续序号的语义不相关性;(ii)在主机入侵检测领域中采用 DCGAN 框架进行小样本数据生成,提出一种在 GAN 较少迭代周期内制造更多次收敛的技巧,根据迭代过程中的生成器与判别器损失函数的接近趋势自动重设初始学习率,从而以较低的计算代价生成更多理想样本;(iii)经过一定次数迭代后,选取当生成器与判别器的损失函数最接近时的若干次输入样本,与原数据集进行混合,通过多种检测方法、数据混合方法来验证其生成效果.

2.1 深度卷积生成对抗网络

经典的生成对抗网络由生成器 G 和判别器 D 组成[8],训练过程如下:首先,定义噪声变量 $p_z(z)$,z 表示噪声数据,通过生成器映射到数据空间形成 $G(z;\theta_g)$,其中 G 是一个多层感知机判别函数,由参数 θ_g 决定;然后,定义第 2 个多层感知机 $D(x;\theta_d)$,θ_d 为参数,其输出一个标量. $D(x)$ 表示 x 的概率,由数据决定,而不是由 p_g 决定. 通过训练判别器 D 给训练数据和生成器生成的样本进行分类,最终是要训练生成器 G 从而最小化 $\log(1-D(G(z)))$.

也就是说,G 和 D 通过下式的目标函数 $V(D,G)$ 展开 2 元极小极大博弈:

$$\min_G \max_D V(D,G) = E_{x\in P_{data}}(\log D(x)) + E_{x\in p_z(z)}(\log(1-D(G(z)))), \quad (1)$$

其中判别器的目标函数是一个交叉熵函数:

$J^{(D)} = -E_{x\in P_{data}}\log D(x)/2 - E_z\log(1-D(G(z)))/2$，$-E_{x\in P_{data}}\log D(x)/2$ 表示判别器判断出 x 是真正来自训练数据的情况，$-E_z\log(1-D(G(z)))/2$ 表示生成器伪造噪音数据的情况. 同理，生成器的目标函数 $J^{(G)}$ 为 $-J^{(D)}$. 最终得到的平衡就是 $J^{(D)}$ 的鞍点.

对判别器 D 而言，它的优化取决于 $D(x) = P_{data}(x)/(P_{data}(x) + P_{model}(x))$，最终要达到的效果是 $D(x)$ 无限接近于 1/2. 由于在平衡状态下，$D(x)$ 的导数值为 0，模型在到达稳定之后便饱和了. 为了解决饱和问题，G 的目标函数可改进为根据伪造的成功率来决定，即 $J^{(G)} = -E_z\log(1-D(G))/2$. 这样，均衡不再由损失决定，在 D 完美之后，G 还可以继续被优化.

在图像生成领域上，A. Alec 等[9] 将卷积神经网络引入到生成模型和判别模型中，形成深度卷积生成对抗网络（Deep Convolutional Generative Adversarial Networks，DCGAN），使得图像生成效果有了较大提升. 其构建要点包括：(i) 将判别器中的池化层用卷积步长代替，在生成器中将卷积层用反卷积代替；(ii) 在生成器和判别器中进行批量归一化；(iii) 深层架构中移除全连接的隐藏层；(iv) 生成器输出层使用 Tanh 作为激活函数，其他层采用 ReLU 激活函数；(v) 判别器所有层用 LeakyReLU 作为激活函数. 在生成模型中通过反向卷积神经网络把噪声数据的特征放大，判别模型对生成的假数据与原始数据融合后由判别模型的卷积神经网络进行逐层降采样学习.

2.2　小样本生成及其优化

2.2.1　**小样本数据生成过程**　利用 DCGAN 生成主机序列异常样本，包括以下 4 个步骤：(i) 数据处理，将主机序列进行抽取、填充和转化为等长向量，随机取一部分异常数据用于生成噪声；(ii) 搭建 DCGAN 模型，其中判别器的结构是一个单层卷积神经网络，卷积层采用 ReLU 激活函数，输出层采用 sigmoid 激活函数；生成器是一个去除池化层的反向卷积神经网络，反卷积层和输出层均采用 ReLU 激活函数. 2 个模型初始学习率根据具体情况分别采用 SGD 和 Adam 优化方法调整生成器和判别器的学习率. (iii) 通过生成数据，预训练判别模型；(iv) 构建训练框架，其核心伪代码如算法 1 所示.

算法 1　生成式对抗网络训练框架

输入：训练数据，完成预训练的 D.

输出：每一步 D 和 G 的损失情况 D_{Loss} 和 G_{Loss}，以及噪声数据.

a) for each epoch，

b) 取 n 个真实训练数据 d_{ata}，

c) 生成器 G 生成 n 个噪声数据 z，

d) 计算数据来自原始数据的概率 P_{data} 和来自生成数据的概率 p_z，

e) 将真实数据 d_{ata} 标记为 1、噪声数据 z 标记为 0，

f) $D_{Loss} = -\frac{1}{n}\left(\sum_{x\in P_{data}}\log D(x) + \sum_{x\in p_z}\log(1-D(G(z)))\right)$，

g) 更新 D 的参数，保持 G 不变，

h) 标记噪声数据为 1，

i) $G_{Loss} = \frac{1}{n}\sum_{x\in p_z}\log(D(G))$，

j) 更新 G 的参数，保持 D 不变，

k) 记录本次迭代的噪声数据 z、D_{Loss} 和 G_{Loss}.

2.2.2　**优化训练过程生成更多样本**　在深度学习诸多优化方法中，Adam 是一种灵活且适用性强、善于处理较稀疏梯度以及非平稳目标的自适应梯度优化方法，同时考虑了梯度的 1 阶矩估计和 2 阶矩估计，经过偏置校正后对学习率形成有范围的动态约束. 假设 m_t、n_t 分别为 t 时刻梯度的 1 阶和 2 阶矩估计，即

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t,$$

$$n_t = \beta_2 n_{t-1} + (1-\beta_2)g_t^2.$$

$\hat{m}_t$ 和 $\hat{n}_t$ 是对 m_t 和 n_t 的校正，校正方法为 $\hat{m}_t = m_t/(1-\beta_1^t)$，近似为对期望的无偏估计，则步长变化为

$$\Delta\theta_t = -\eta\hat{m}_t/(\sqrt{\hat{n}_t} + \varepsilon),$$

其中 ε 为使分母不等于 0 的极小正值.

但是，在利用 GAN 生成样本时需考虑的问题是，由于采用自适应梯度的优化方法，梯度在训练过程中递减，所以梯度在接近收敛状态时接近消失，这容易造成模式崩溃. 为此，本文提出一种重设 Adam 梯度的技巧，从而对训练过程进行自动干预，以构造更多的平衡状态. 具体方法是：在训练过程中设置一种机制，当模型接近收敛且呈现模式崩溃趋势时，即当 t 满足

$$|G_{Loss_t} - D_{Loss_t}| < \gamma D_{Loss_t}, G_{Loss_t} > \frac{1}{k}\sum_{i=t-k}^{t-1} G_{Loss_i}$$

时，重设梯度优化参数，重设 $\beta_1^t \leftarrow \beta_1^0$，$\beta_2^t \leftarrow \beta_2^0$，然后继续训练，从而控制损失，提高收敛速度. 其中 γ 为较小的百分比，k 为较小的正整数.

2.3　构建检测序列

主机序列领域知名数据集包括 DARPA、UMN、ADFA 等，由于 DARPA、UMN 数据集已有 20 年以上

历史,且内容不足以反映攻击行为的复杂性,因此本文使用 ADFA-LD[10] 主机序列数据集展开实验. 该数据集由澳大利亚国防学院于 2012 年发布,被广泛应用于入侵检测类产品的测试. 其概括了当前市面上最常见的 Linux 系统服务器,这些服务器提供文件共享、数据库服务、远程连接和 web 服务器的功能,同时也有一些小型的残余漏洞. 在制作这个数据集时,配置人员已经仔细地考虑了渗透测试人员和黑客通常采用的方法,将不同的系统调用以不同的正整数表示,并对系统调用进行了特征化,对攻击类型进行了标注.

在检测序列的构造上,本文的方法首先对该数据集采用一种 Seq2Image 的数据处理方法,将不同长度的序列填充为固定长度后再归一化,转换为 $m \times m$ 长度的张量,通过热图的方式进行可视化; 然后将该数据输入按上述方法构建的深层卷积生成对抗网络框架,经过若干次迭代和优化后,取在 D 和 G 损失函数值最接近时的迭代样本作为输出; 最后将输出样本还原到 $[1,340]$ 的样本空间,和原始数据集中的其他异常数据进行混合,从而构成下一步检测所需的序列.

2.4 异常检测

由于本文拟通过多种检测方法对生成样本的效果进行验证,故先对主机序列异常检测领域的常用方法进行简要回顾.

使用 n-Gram 进行序列化特征建模是最传统的技术,最初见于 S. Forrest 等[11] 提出的 STIDE 模型,基本思路是将系统调用看作词语,把调用序列看作短语,设 k 为序列长度,则窗口大小为 $k+1$,在滑动窗口时用数据库记录每个词的后续序列集. 此后,部分学者对其进行了扩展,如 B. Subba 等[12] 为了解决爆发式增长的特征向量问题,提出一种轻量级的模型,只考虑频率大于阈值的 n-Gram 词语特征,在 ADFA-LD 数据集表现出较高的准确率和较低的误报率,且产生较低的计算负担. 近年来,G. Serpen 等[13] 和 E. Aghaei 等[14] 对 ADFA 数据集有较多基于 n-Gram 滑窗的研究,如将 ADFA 数据集分为1 个正类和 6 个攻击类别,通过滑窗和主成分分析,对固定长度的系统调用序列原始数据提取特征进行训练,标准化后以特征向量的形式对序列进行表示,然后用 KNN、随机森林、RBF 神经网络等方法来完成分类.

在 n-Gram 基础上,基于隐马尔可夫模型的技术是 较为热门且效果较好的一种检测方法. 但是,基于隐马尔可夫模型的方法计算代价较高. C. Warrender 等[15] 比较了 4 种基于 n-Gram 的检测方法,结论是基于 HMM 方法可以达到较高准确率,但该方法代价较大,且检测结果在不同测试集上的效果差别较大,算法区别不明显. Gao Debin 等[16] 提出一种新颖的基于 HMM 的检测方法,其在提高准确率和降低能耗方面均有所改善.

由于主机入侵检测系统越来越多部署于物联网环境,所以往往对计算代价更为敏感,学者开始考虑代价较低的特征提取办法. 如对序列提取基于频率的特征,其中以基于 TF-IDF 的主机序列异常检测方法为代表,Liao Yihua 等[17] 提出将系统调用进行向量化计算,先用 TF-IDF 对序列评分,并在此基础上,对 SVM、kNN 等算法进行分类. Zhang Zonghua 等[18] 提出2 种降低误报率的方法. R. Vijayanand 等[19] 针对物联网和 IDS 需实施监控数据流量的环境,提出一种计算负担较小的基于遗传算法和互信息的混合特征选择技术,然后通过 SVM 进行分类.

除上述异常检测方法外,本文针对主机序列命令 one-hot 表示的稀疏性特点,提出采用基于 Lasso-kNN 的检测方法,即采用 Lasso 回归的特征提取方法,然后使用 kNN 分类进行检测. Lasso 回归是在普通线性回归模型基础上增加 L_1 范数作为惩罚约束,当 λ 充分大时,可以把某些待估计系数准确地收缩到 0,从而起到对稀疏数据进行有效降维的效果. 其损失函数为

$$w^* = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{i=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|,$$

其中 p 为模型中的系数个数. 在求解过程中,需要先通过多次迭代对 λ 进行训练,获得最佳的 λ,进而再次进行拟合.

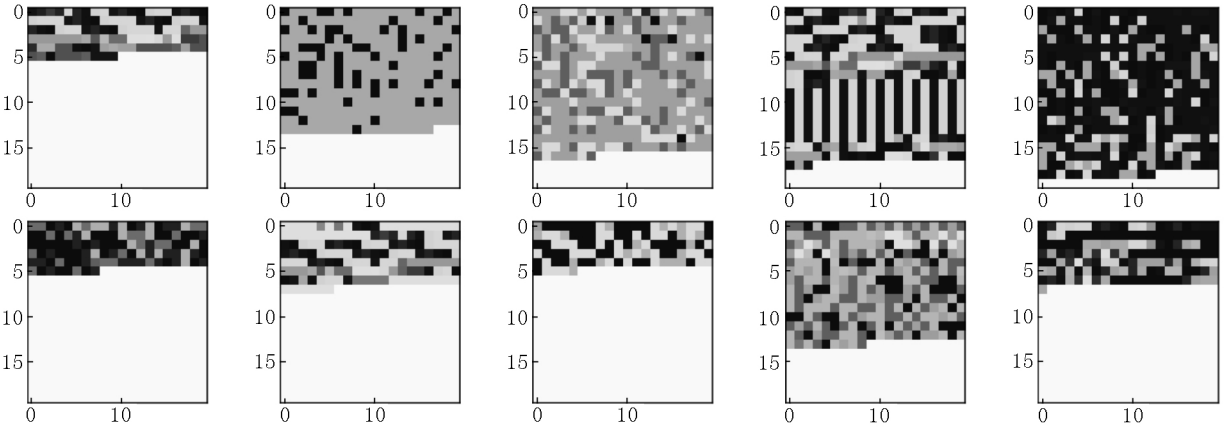
3 实验及分析

3.1 数据处理

ADFA 数据集正常与异常样本分别有 5 206 和 719 条,比例接近 7∶1,每条样本代表一串序列,序列长度介于几十到几千之间,序列中的命令以 1 ~ 340 之间的正整数编号来表示. 数据的规模即序列的长度,是根据异常检测场景而定,对于本数据集而言,序列长短差别较大,取长度不大于 400 的序列,接近数据集中序列长度的中位数,然后将长度不足 400

的部分用 $n+1$(n 为异常序列维度) 即 341 填充，取值映射到 [0,1] 内，形成一批 20×20 的张量.

通过 Seg2Image 的数据处理方法，将正常序列和异常序列转化为 20×20 的热图(见图 2) ，其中第 1 行图像由正常序列转换而成，第 2 行图像由异常序列转化而成. 使用的热图类型为 gist-ncar，因为它显示了丰富且层次分明的色彩变化，适用于表现不同的主机命令.



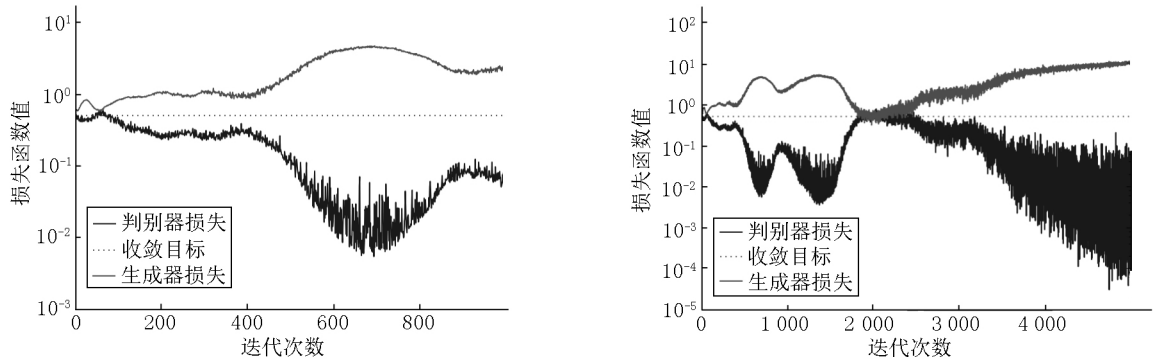
注: 第 1 行由正常序列转换而成，第 2 行由异常序列转换而成.

图 2　将 ADFA-LD 数据集的正常序列和异常序列转换为 20×20 像素的图像

3.2　利用深度卷积对抗网络生成异常样本

先设计生成对抗网络框架. 在生成器 G 中，使用激活函数为 ReLU 的全连接层制造噪音数据，预训练所生成的模型将由 GAN 的首次迭代的判别器进行判别. 判别器 D 设计为一个 CNN，池化层采用 ReLU 函数作为激活函数，卷积层采用 sigmoid 作为激活函数.

对于训练部分，先输入一定量的异常数据生成噪声，对 G 进行预训练，然后多次迭代 D 和 G 进行对抗训练，并记录在每次迭代时它们的损失函数. 在训练方面，生成器采用 SGD 优化方法，初始学习率设置为0.08，判别器采用 Adam 优化，初始学习率设置为2.2×10^{-5}. 图 3 显示了在 1 000 次和 5 000 次迭代中 D 和 G 的损失函数变化情况，在此实验过程中，当第 1 846 次迭代时 G 和 D 的损失函数值最接近，约为0.53，接近理论上的收敛目标值 0.5.



(a) 1 000 次迭代训练损失　　(b) 5 000 次迭代训练损失

图 3　通过损失函数表示的生成器与判别器迭代过程

图 4 显示了在表现最佳的某次迭代中，生成器生成的 10 条异常样本转化成的 20×20 图像，由于 x 轴和 y 轴在图像表示中没有意义，故在此图中将其隐去.

3.3　提高收敛频率

图 3 所示训练过程通过控制初始学习率将训练损失控制在一定范围，但训练过程生成器与判别器的收敛密度较小，接近收敛状态后随即散开，收敛情况不受控制. 为进一步加快收敛进程，在较少迭代次数内获得更多平衡状态的生成序列，笔者在训练阶段中根据生成器和判别器的损失值，构建一套重设学习率初始值的机制. 重设条件为: 当 2 者接近且在趋势上呈发散状态时，即当某次迭代时，生成器损失值超出判别器损失值的 $\pm\gamma$(%) ，且 k 次迭代移动平均线呈上升趋势，则重设生成器和判别器的优化方法和参数，迅速增加学习率. 在训练过程中使用 Adam优化方法自动调整 2 者学习率，当达到重设条件时重设初始学习率，本实验重设的初始学习率为 2×10^{-5}，参数 β_1 和 β_2 均设为 0.9.

图 5 显示了当 k 取值为 3,而 γ 取值分别为 5%、10%、15%、20%时重设学习率的 1 000 次迭代训练损失收敛情况,可见 γ 的合理取值能在一定程度上增加收敛密度,从而在较少迭代次数内获得更多收敛状态的生成样本. 在本实验中,在 1 000 次迭代内,本文方法将收敛频率从 1 次提高到 5 次以上,这意味着可将生成样本的计算成本压缩至约 20%.

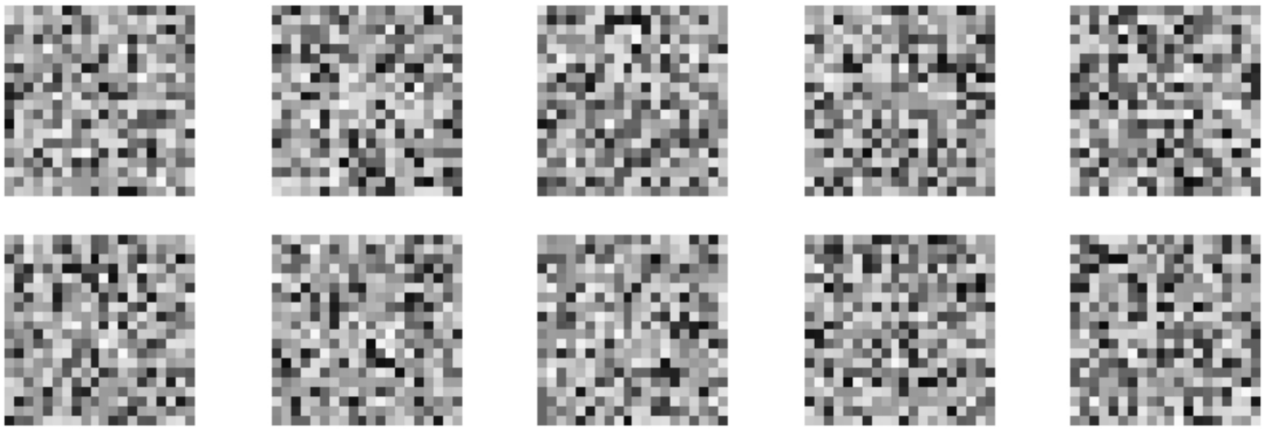
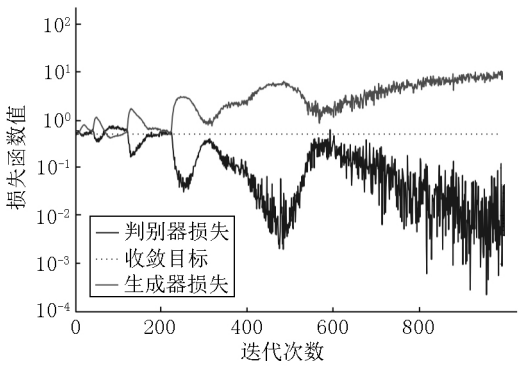
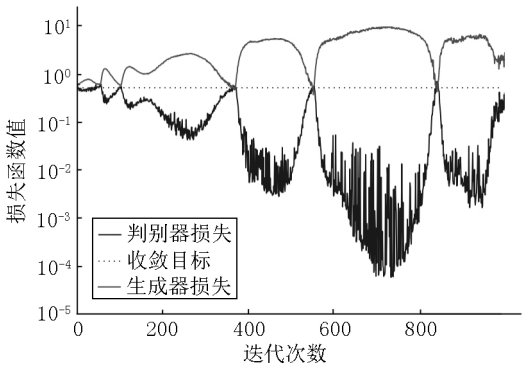


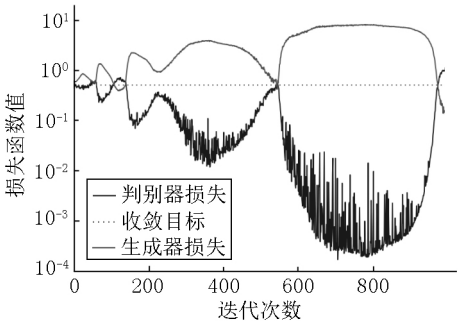
图 4 表现最佳的 1 次迭代所生成的 10 条异常序列



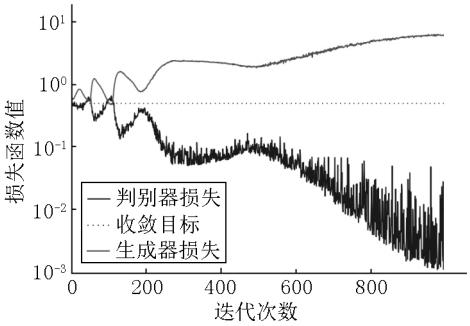
(a) 1 000 次迭代训练损失($\gamma=0.05$)



(b) 1 000 次迭代训练损失($\gamma=0.10$)



(c) 1 000 次迭代训练损失($\gamma=0.15$)



(d) 1 000 次迭代训练损失($\gamma=0.20$)

图 5 在训练过程中重设学习率,增加收敛次数

3.4 检测结果比较

为使实验起到平衡样本的作用,在训练数据中配置 960 条正常样本,以及 160 条原 ADFA-LD 数据集的异常样本、800 条生成的异常样本,2 者之比为 1∶5,训练数据与测试数据之比为 4∶1,在测试数据中生成样本按照与训练集同样比例配置. 实验采用基于 STIDE、Lasso-kNN、TF-IDF-kNN 这 3 种不同的异常检测算法,对添加生成样本前后的检测结果进行对比. 生成异常样本的检测指标为 TPR 和 FPR. TPR 代表异常样本的检出率(又称召回率),即异常样本被正确识别的数量占总样本的比例. FPR 代表误报率,即将正常样本被当作异常样本的数量占总样本数的比例.

3.4.1 与不平衡数据的检测结果对比 为了更加全面地反映生成样本的数据平衡能力,本实验在 3 种不同的数据配比方法下检测数据平衡结果:(i) 在上述不平衡数据集中加入生成样本后,重新划分为 80% 训练集和 20% 测试集,其中测试数据中包含用

于生成异常样本的种子数据; (ii) 在(i) 情形下，测试数据中不包含种子样本; (iii) 仅在训练集上加入生成样本，而测试集包含种子样本，不加入生成样本.

实验通过基于 STIDE、Lasso-kNN、TF-IDF-kNN 的检测方法对上述 3 种情形进行异常检测，实验结果(见表 1) 表明: 在情形(i)、(ii) 下，使用改进的 DCGAN 生成的异常样本进行数据平衡能有效改善检出效果，尤其是在基于 n-Gram 的 STIDE 方法下，本文方法的样本平衡明显提高了异常样本检出率和降低了误报率，而在基于 Lasso-kNN 和 TF-IDF-kNN 特征提取的异常检测方法下，本文方法也能有效提高异常样本的检出率，且使误报率维持在较低水平. 对于情形(iii)，本文方法仅对部分检测方法有效.

表 1　在不同数据分配情形下采用 3 种异常检测方法验证本方法的样本平衡效果　%

数据分配方式	异常检测方法	平衡前(6∶1)		平衡后(6∶6)	
		TPR	FPR	TPR	FPR
情形(i)	STIDE	60.0	15.0	85.6	1.7
	Lasso-kNN	92.5	2.5	98.5	2.5
	TF-IDF-kNN	87.5	5.8	97.9	6.3
情形(ii)	STIDE	55.0	15.0	83.3	1.7
	Lasso-kNN	92.5	2.5	98.8	2.5
	TF-IDF-kNN	87.5	5.8	97.9	6.3
情形(iii)	STIDE	57.8	16.6	57.8	16.6
	Lasso-kNN	89.1	2.2	86.7	2.2
	TF-IDF-kNN	89.1	2.2	90.4	2.2

3.4.2　与其他数据平衡方法检测结果对比　基于上述 3 种异常检测方法，将本文所述的数据平衡方法与其他数据平衡方法生成的异常样本检测结果进行对比. 实验使用的降采样、随机过采样方法分别是随机减少正常样本和随机复制异常样本，使2者比例达到 1∶1. 此外，也与基于未改进的 DCGAN 生成的异常数据进行检测结果对比.

表 2 结果显示: 在 3 种异常检测方法下，过采样方法的数据平衡效果优于降采样. 其中本文方法的数据平衡方法在基于 n-Gram 的 STIDE 方法下有较好的平衡效果，能大幅提升检测率并降低误报率; 在基于特征提取的 Lasso-kNN、TF-IDF-kNN 方法上，本文方法与随机过采样的数据平衡效果相近，其中经过训练优化的 DCGAN 方法在 TF-IDF-kNN 异常检测方法下表现略好.

令人意外的是，当使用 Lasso-kNN 和 TF-IDF-kNN 方法进行异常样本检测时，采用随机过采样方法平衡数据能使异常数据检测率和误报率达到较高水平. 笔者分析认为，这可能是因为与训练数据重复的异常样本造成过拟合因而大大改善了测试数据的检测结果，而该方法本身并没有真正生成新的异常数据. 在这个环节的对比实验中，笔者并未采用 SMOTE 方法，如第 1 节所述，在本数据集中系统调用序列虽然表示为正整数，但语义距离与表示数值大小并无关联，不适合用 SMOTE 方法生成异常，因此笔者使用随机过采样的平衡方法替代 SMOTE，可认为是 SMOTE 在随机选择数据样本时的一种极端情况.

上述实验结果说明，本文提供的方法较其他数据平衡方法，能较好地提升检测结果，尤其是对于改善基于 STIDE 的检测方法的实验结果上，通过 DCGAN 生成异常样本能有效提升异常数据的检测率并降低误报率，能在较低的数据量下大幅提高 STIDE 方法对异常数据检测的准确率.

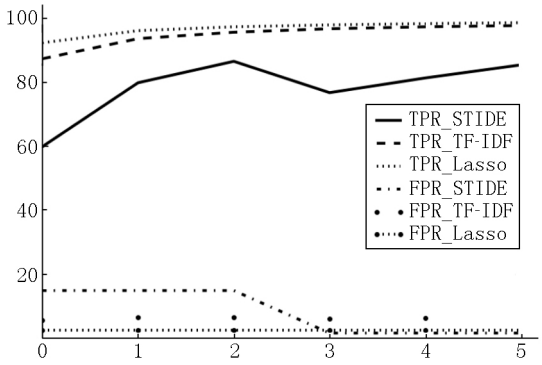
表 2　不同数据平衡方法的平衡效果比较　%

数据平衡方法	STIDE		Lasso-kNN		TF-IDF-kNN	
	TPR 提升	FPR 降低	TPR 提升	FPR 降低	TPR 提升	FPR 降低
训练优化的 DCGAN	21.7	16.2	12.5	−0.4	10.0	1.7
DCGAN	21.6	16.3	12.5	−0.4	10.0	1.3
随机过采样	−2.0	0	13.3	0	9.2	2.1
降采样	−20.0	2.5	−5.0	−7.5	5.0	−6.7

3.4.3　不同数据平衡程度下的检测结果　为了检测不同数据平衡程度下的检测结果，在异常数据与正常数据之比为 1∶6 的基础数据集上逐步增加异常生成样本的比例，直到数据平衡，然后利用上节所述的 3 种异常检测方法进行检测，实验结果如表 3 和图 6 所示. 结果显示: 随着生成样本的增加，异常样本的检测率有明显提升，且误报率维持在较低水平. 尤其是 STIDE 方法，在数据集中配置生成的异常样本，不仅提升了检测率，也将误报率从 15.0% 降低到1.7%. 此外，在增加生成样本的过程中，对已检出的异常样本进行跟踪，结果是在 3 种检测方法下，新增加的生成样本均能被检测出，且已检测出的异常样本在增加生成样本后均无漏检.

表 3 在不同生成样本比例下的异常检测结果 %

增加生成样本比例	TPR			FPR		
	STIDE	TF-IDF-kNN	Lasso-kNN	STIDE	TF-IDF-kNN	Lasso-kNN
不增加	60.0	87.5	92.5	15.0	5.8	2.5
增加 1 倍	80.0	93.8	96.3	15.0	6.7	2.5
增加 2 倍	86.7	95.8	97.5	15.0	6.7	2.5
增加 3 倍	76.9	96.9	98.1	1.7	6.2	2.5
增加 4 倍	81.5	97.5	98.5	1.7	6.3	2.5
增加 5 倍	85.6	97.9	98.8	1.7	6.3	2.5



注: 横轴表示增加生成样本的倍数.

图 6 在数据平衡过程中异常样本检测率和误报率变化情况

4 结束语

本实验在主机异常检测领域中,利用改进的深层卷积生成对抗网络构造 ADFA-LD 数据集的异常数据进行数据过采样平衡,获得了较为显著的结果,通过多种检测方法实验验证了基于深层卷积生成对抗网络进行生成异常对抗样本的有效性,尤其是对于基于 STIDE 的检测方法,通过数据平衡较大提升了异常样本检测率. 此外,本文提出一种针对生成对抗网络的训练优化策略,能在较少迭代次数内生成更多理想的生成样本,从而大幅降低训练成本,该方法在其他数据集上的应用效果尚未可知,后续将进一步对该策略展开泛化研究.

5 参考文献

[1] Zachary G,Schwartz S. Data preprocessing and feature selection for an intrusion detection system dataset [EB/OL]. [2019-12-16]. https://www.researchgate.net/publication/331730342_Data_preprocessing_and_feature_selection_for_machine_learning_intrusion_detection_systems.

[2] Shekarforoush S,Green R C,Dyer R,et al. Classifying commit messages: a case study in resampling techniques [EB/OL]. [2019-12-16]. http://ieeexplore.ieee.org/iel7/7958416/7965814/07965999.pdf.

[3] Nitesh V,Kevin W,Lawrence O,et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research,2012,16: 321-357.

[4] Liu Kaijian,Fan Zhen,Liu Meiqin,et al. Hybrid intrusion detection method based on *K*-means and CNN for smart home [EB/OL]. [2019-12-16]. https://www.researchgate.net/publication/332378390_Hybrid_Intrusion_Detection_Method_Based_on_K-Means_and_CNN_for_Smart_Home.

[5] Promper C,Engel D,Green R C. Anomaly detection in smart grids with imbalanced data methods [EB/OL]. [2019-12-16]. http://www.en-trust.at/papers/promper17a.pdf.

[6] Lee J,Park K. GAN-based imbalanced data intrusion detection system [EB/OL]. [2019-12-16]. https://www.researchgate.net/publication/337169111_GAN-based_imbalanced_data_intrusion_detection_system.

[7] Salem M,Taheri S,Yuan J S,et al. Anomaly generation using generative adversarial networks in host based intrusion detection [EB/OL]. [2019-12-16]. http://arxiv.org/abs/1812.04697.

[8] Goodfellow I,Pougetabadie J,Mirza M,et al. Generative adversarial nets [EB/OL]. [2019-12-16]. http://www.iro.umontreal.ca/~lisa/publications2/index.php/publications/show/808.

[9] Radford A,Metz L,Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks [EB/OL]. [2019-12-16]. https://arxiv.org/abs/1511.06434v2.

[10] Creech G,Hu Jiankun. A semantic approach to host-based intrusion detection systems using contiguous and discontiguous system call patterns [J]. IEEE Transactions on Computers,2014,63(4): 807-819.

[11] Forrest S,Hofmeyr S,Somayaji A,et al. A sense of self for Unix processes [M]. New York: IEEE Symposium on Security and Privacy,1996: 120-128.

[12] Subba B,Biswas S,Karmakar S. Host based intrusion detection system using frequency analysis of *N*-gram term [EB/OL]. [2019-12-16]. https://www.researchgate.net/publication/322218531_Host_based_intrusion_detection_system_using_frequency_analysis_of_n-gram_terms.

[13] Serpen G,Aghaei E. Host-based misuse intrusion detection using PCA feature extraction and kNN classification algo-

rithms [J]. Intelligent Data Analysis,2018,22(5):1101-1114.

[14] Aghaei E,Serpen G. Host-based anomaly detection using Eigentraces feature extraction and one-class classification on system call trace data [R/OL]. [2019-12-16]. https://arxiv.org/abs/1911.11284.

[15] Warrender C,Forrest S,Pearlmutter B. Detecting intrusions using system calls: alternative data models [EB/OL]. [2019-12-16]. https://ieeexplore.ieee.org/document/766910.

[16] Gao Debin,Reiter M K,Song D. Behavioral distance measurement using hidden markov models [EB/OL]. [2019-12-16]. https://www.researchgate.net/publication/221427551_Behavioral_Distance_Measurement_Using_Hidden_Markov_Models.

[17] Liao Yihua,Vemuri V R. Use of K-nearest neighbor classifier for intrusion detection [J]. Computers and Security,2002,21(5):439-448.

[18] Zhang Zonghua,Shen Hong. Application of online-training SVMs for real-time intrusion detection with different considerations [J]. Computer Communications,2005,28(12):1428-1442.

[19] Vijayanand R,Devaraj D,Kannapiran B,et al. A novel intrusion detection system for wireless mesh network with hybrid feature selection technique based on GA and MI [J]. Journal of Intelligent and Fuzzy Systems,2018,34(3):1243-1250.

[20] Bridges R A,Glass-Vanderlan T R,Iannacone M D,et al. A Survey of intrusion detection systems leveraging host data [J]. ACM Computeng Surveys,2020,52(6):1-35.

[21] Teng Shaohua,Wu Naiqi,Zhu Haibin,et al. SVM-DT-based adaptive and collaborative intrusion detection [J]. IEEE/CAA Journal of Automatica Sinica,2018,5(1):108-118.

[22] Deshpande P,Sharma S C,Peddoju S K,et al. HIDS: a host based intrusion detection system for cloud computing environment [J]. International Journal of System Assurance Engineering and Management,2018,9(3):567.

[23] Kashyap A,Kumar G S,Jangir S,et al. IHIDS: introspection-based hybrid intrusion detection system in cloud environment [EB/OL]. [2019-12-16]. https://ieeexplore.ieee.org/iel7/8119306/8125802/08125921.pdf.

[24] Creech G,Hu Jiankun. A semantic approach to host-based intrusion detection systems using contiguous and discontiguous system call patterns [J]. IEEE Transactions on Computers,2014,63(4):807-819.

[25] Xie Miao,Hu Jiankun,Slay J,et al. Evaluating host-based anomaly detection systems: application of the one-class SVM algorithm to ADFA-LD [EB/OL]. [2019-12-16]. https://www.researchgate.net/publication/287318600_Evaluating_host-based_anomaly_detection_systems_Application_of_the_one-class_SVM_algorithm_to_ADFA-LD.

[26] Msika S,Quintero A,Khomh F. SIGMA: strengthening IDS with GAN and Metaheuristics attacks [EB/OL]. [2019-12-18]. https://arxiv.org/abs/1912.09303.

[27] 滕少华,孔棱睿. 基于生成式对抗网络的中文字体风格迁移 [J]. 计算机应用研究,2019,36(10):3164-3167.

[28] 卢逸君. 一种主机序列入侵检测方法: 中国,2019105964097 [P]. 2019-10-15.

The Generation of Minority Sample Data and Its Application in Abnormal Detection

LU Yijun[1,2],TENG Shaohua[1*]

(1. College of Computer,Guangdong University of Technology,Guangzhou Guangdong 510006,China;
2. Guangdong Information Technology Security Evaluation Center,Guangzhou Guangdong 510095,China)

Abstract: In the application of unbalanced data,the small number of negative samples (abnormal data) can be an important reason for low detection rate,as in the field of host based intrusion detection,the gap of sample size for majority class and minority class can lead to poor detection result. To solve this problem,the deep convolutional generative adversarial networks(DCGAN) are improved in the paper,making it easier to converge and generate more ideal samples,which introduces improved DCGAN to the intrusion detection evaluation data set ADFA-LD and generates more abnormal samples to make the data set more balanced. Finally,a variety of abnormal detection methods are used in the paper to observe the effect of this data - balancing method. The result shows that newly generated abnormal samples can all be detected,without missing any detected abnormal sample,which leads to higher detection rate and lower false positive rate. Therefore,it is concluded that this data generation method can effectively alleviate some data imbalance problems in practice.

Key words: convolutional neural networks; generative adversarial networks; sample generation; host-based intrusion detection; neural network

(责任编辑: 冉小晓)