DOI:10.13718/j.cnki.xsxb.2021.07.010

# 基于对抗性自编码器的入侵检测算法①

白洁仙[1], 剧雷鸣[2]

1. 山西工商学院 计算机信息工程学院, 太原 030006;
2. 南阳理工学院 软件学院, 河南 南阳 473000

**摘要:** 针对当前无监督学习的入侵检测算法准确度低、误报率高以及有监督学习算法所需训练样本标记成本高的问题, 提出一种基于对抗性自编码器的入侵检测算法. 这是一种半监督学习算法, 仅需要训练数据集中少量标记数据进行训练, 并在训练数据集中支持未标记数据, 从而提高性能. 首先, 自编码器通过提取重要特征作为潜在变量来降低输入数据的维数; 其次, 利用生成对抗网络使自编码器的潜在变量遵循任意分布以进行正则化; 最后, 利用标记数据的交叉熵损失来实现半监督学习的分类. 实验结果表明: 相较于其他算法, 本文所提算法对少量标记的数据集检测具有一定的优势, 在实现高准确度、低误报率的同时, 降低对标记数据的需求.

**关 键 词:** 入侵检测算法; 自编码器; 生成对抗网络; 半监督学习

**中图分类号:** TP393 **文献标志码:** A **文章编号:** 1000-5471(2021)07-0077-07

随着计算机网络与日常生活的联系越来越紧密, 网络安全问题也越来越严重, 如数据丢失、拒绝服务、非法访问以及黑客攻击等等[1-2]. 其中, 拒绝服务(Denial of Service, DoS)攻击通过引入不需要的流量来拒绝或阻止网络上的合法用户资源, 恶意软件则使用恶意软件破坏系统[3]. 因此, 作为主动安全防范措施的入侵检测系统(Intrusion Detection System, IDS)被用来监视计算机和网络系统的活动, 拦截各种网络攻击和入侵, 成为网络安全领域研究的热点内容.

IDS 方法大致可以分为基于签名或基于异常两类[4]. 基于签名的入侵检测系统[5]对已知攻击进行分析, 提取特征和模式, 并以此为基础建立数据库. 该方法的优点是检测率高, 对已知攻击的误判率低, 但是无法检测未知和新的攻击. 基于异常的 IDS[6]将新数据与正常用户行为模型进行比较, 将与此模型出现显著偏差的流量标记为异常, 这类方法的优势在于可以检测到未知的和不可预见的新攻击. 由于异常检测在检测新的攻击方面比签名检测更具优势, 因此受到众多研究人员的关注. 目前, 异常的 IDS 可以利用机器学习来建立基于网络流量特征的分类器模型. 随着深度学习的出现和发展, 手工定义特征的任务被可训练的多层网络所取代, 从而能够更好地解决当前入侵检测算法面临的准确度低和误报率高的问题, 因此各类深度学习方法广泛应用于 IDS 领域中. Potluri 等[7]将卷积神经网络(CNN)应用于入侵检测机制中, 能够有效地识别新的攻击. Xu 等[8]提出一种基于深度学习和转移学习的入侵检测方法, 将入侵检测问题转化为图像识别问题, 该方法具有更好的泛化性能, 能够更有效地检测新的入侵方法. 但是, 一些基于深度学习的入侵检测系统在实现高检测率的同时, 往往伴随着较高的误报率, 这可能会大大降低入侵检测系统的整体效能. 为了解决这一问题, Wang 等[9]采用了卷积神经网络(CNN)和长期短期记忆网络(LSTM)来进行特征提取, 通过捕获时空特征来实现有效的特征学习用于降低误报率. Wu 等[10]采用层次化的入侵检测模型, 分别利用 CNN 和循环神经网络(RNN)以逐渐增加粒度的方式同步学习输入的数据, 通过考虑网络流

① 收稿日期: 2020-07-24
基金项目: 全国教育科学规划教育部重点课题(QN020515).
作者简介: 白洁仙, 硕士, 讲师, 主要从事计算机技术研究.

量数据中时空特征的存在来实现检测模型的高检测能力和低误报率. 虽然准确率和误报率一直是IDS研究的重点，但是实时性和检测效率也是一项很重要的指标. 在深度学习中，相较于其他神经网络，自编码器能够有效地降低特征维度，并容易与其他模型相结合，在提高检测精度的同时，大大降低模型的训练时间. Alqatf等[11]提出了一种基于堆叠自编码器的深度学习框架，通过大量的样本训练来识别攻击流量的特征. Mirza等[12]首先利用自编码器对数据进行降维，然后使用LSTM网络来提取样本流量的时域特征，从而有效应对不可预见和不可预测的网络攻击. 但是，这类方法在利用自编码器压缩原始数据时，不可避免地要丢失掉信息. 而且许多有监督的深度学习方法需要使用带有正确标签注释的训练数据进行训练，并且监督学习需要大量标记数据才能实现高精度. 但是，标记数据是一项昂贵的任务，工作人员需要检查数据，然后对其进行分类并使用适当的标签进行注释. 此外，受IDS约束的网络流量趋势每天都在变化，且会继续产生新的攻击. 因此，标记工作需要进行多次，并且难以保留足够数量的标签数据.

针对上述存在的问题，本文创新性地提出了一种基于对抗性自编码器的入侵检测算法，该方法将对抗的思想融入到自动编码器中，通过使用对抗学习替代相对熵的使用来提高检测准确性，同时减少了计算量. 而且作为半监督学习算法，仅使用训练数据集中的少量标记数据来训练学习分类器，并在训练数据集中支持未标记数据，因而有效地减少了昂贵的人工任务.

## 1 相关内容

### 1.1 自编码器

自编码器(Auto-encoder，AE)是一种无监督学习的神经网络，它使用了反向传播算法，并让输出数据$X'$等于输入数据$X$. AE经常被用于高维数据的降维，训练数据随机生成或者作为强大的特征检测器应用在深度神经网络的预训练中.

AE由编码器和解码器两个神经网络组成. 首先，编码器将原始数据$X$映射到潜在空间中，该步减小了输入数据的维数，其输出为压缩后的中间层潜在向量$z$. 其次，解码器扩展潜在向量$z$的维数来重构原始输入数据. 为了使输入和输出数据之间的损失最小，需要训练与编码器、解码器相关的神经网络参数. AE的目的是在保持数据重要特征的同时，降低输入数据的维数. 只提取数据的重要特征和结构，并将其作为隐藏中间层的潜在变量$z$保存在低维中. 编码器和解码器的定义如下：

$$z = f_\theta(x) = s(Wx + b) \tag{1}$$

$$x' = g_\theta(z) = s_r(W'z + c) \tag{2}$$

式(1)、式(2)中：$x$是输入数据，$x'$是从潜在变量重构的数据，$f_\theta$和$g_\theta$分别表示编码函数和解码函数，$s$和$s_r$是对应的激活函数. 自编码器通过训练来优化参数权重($W$和$W'$)和偏差($b$和$c$)，以最小化重建误差. 神经网络中的中间层$z$是一个潜在变量，它以比输入数据小的维度捕获输入数据的基本特征. 本文方法中使用此潜在变量将正常流量和异常流量分类. 此外，还可以通过假定此潜在变量$z$的任意先验分布对它们进行分类.

### 1.2 生成对抗网络

生成对抗网络(Generative Adversarial Network，GAN)由生成器和鉴别器两个神经网络模型组成. 生成器G尝试捕获训练数据分布，经过训练可以生成类似于真实训练数据($X_{real}$)的生成数据($X_{fake}$)，即生成器G的训练过程是使鉴别器D犯错的概率最大化. 另一方面，鉴别器D试图正确判断对象数据是训练器数据$X_{real}$还是生成器G生成的数据$X_{fake}$，它通过最大化能够正确区分样本来自训练数据而非由生成器G的概率来训练.

GAN在对抗竞争中交替训练两个模型，其最终结果是生成器G能够生成逼近真样本的假样本，且生成的$X_{fake}$与训练数据$X_{real}$的分布相匹配，鉴别器D的输出概率为50%，即判别器无法判别样本的真假. 因此，GAN的目标函数定义为

$$\min_G \max_D V(D, G) = E_{x\sim p_{data}}[\log D(x)] + E_{z\sim p(z)}[\log(1 - D(G(z)))] \tag{3}$$

式(3)中：$G$表示发生器，$G(z)$表示以$z$为输入的发生器生成的数据；$D$是鉴别器，$D(x)$表示对输入$x$进行判别，当输入为真实数据时输出为1，否则为0；$p_{data}$表示实际数据分布，$p(z)$为预定义的先验输入噪声

分布. GAN 能够自动学习输入数据中的规律性或模式，从而可以使用该模型来生成基于原始样本的新样本. 本文使用 GAN 是为了保证 AE 中的潜在变量在对抗性自动编码器模型中遵循任意分布.

## 2　半监督的入侵检测算法

### 2.1　对抗性自编码器

本文提出的基于对抗性自编码器的入侵检测算法是使用自编码器(AE)和生成对抗网络(GAN)作为关键的构成部分. 首先，AE 通过提取和保留重要特征作为潜变量 $z$ 来降低输入数据的维数；其次，利用 GAN 使 AE 的潜变量 $z$ 遵循任意分布进行正则化. AE 的潜在变量 $z$ 由 GAN 的判别器进行正则化，以使任意先验 $p(z)$ 与潜在变量 $z$ 的聚合后验 $q(z)$ 相匹配. 假定 $x$ 为输入向量，$z$ 为 AE 的潜在变量，$p(z)$ 是施加在潜变量上的先验分布，$q(z \mid x)$ 是编码分布，$p(z \mid x)$ 是解码分布，则可以利用自编码器的编码分布对 AE 的潜在变量定义聚合后验分布.

$$q(z) = \int_x q(z \mid x) p_d(x) \mathrm{d}x \tag{4}$$

式(4) 中 $p_d(x)$ 为数据分布，随着学习的进行，对抗性自编码器可以将任意先验 $p(z)$ 与潜在变量 $z$ 的聚合后验 $q(z)$ 相匹配，潜在变量 $z$ 遵循先验分布，即 AE 编码器成为 GAN 生成器. 因此，AE 的潜在变量 $z$ 被视为生成器生成的数据. 鉴别器从遵循 $p(z)$ 分布的发生器生成样本中区分遵循 $q(z)$ 分布的 AE 生成的潜在变量 $z$，获得的输入数据来自任意先验 $p(z)$ 样本的概率 $d$. 对抗网络和 AE 采用随机梯度下降(SGD)方法联合训练.

### 2.2　半监督学习的对抗性自编码器

监督学习需要大量数据来构建一个好的分类器，且带有正确标签注释的训练数据越多，有监督机器学习分类器的性能越好. 但是，获取大量的标记训练数据是一个成本昂贵的过程. 半监督学习只需要训练集中的一小部分标记数据和大量的未标记数据来训练分类器就可以获得很好的分类效果. 这是选择半监督学习的基本动机. 假设输入数据由来自分类分布的潜在变量 $z_1$ 和来自高斯分布的潜在变量 $z_2$ 生成，其中潜在变量 $z_1$ 保存代表类信息的特征(如正常或攻击)，而潜在变量 $z_2$ 保存其他特征，则基于半监督学习的对抗性自编码器利用两对生成模型和判别器对这两个潜在变量 $z_1$ 和 $z_2$ 进行正则化. 使用分类分布 $cat(z_1)$用于描述潜在变量 $z_1$，目标是分离和学习类信息.

分类分布是指采用与类数目相同的独热编码值数目的分布，本文所提模型中对应正常和攻击 2 个类目. 潜在变量 $z_1$ 可由鉴别器学习以保持独热编码，而且与分类分布相对应的潜在变量 $z_1$ 被设计用来记录与输入数据相关联的标签. 由于潜在变量 $z_1$ 保存了类信息，因此可以参考编码器估计的 $z_1$ 值来进行分类. 潜在变量 $z_2$ 则使用高斯分布 $N(z_2|0, 1)$来描述. 本文采用潜在变量 $z_2$ 来保存除类信息以外的详细特征，原因是认为正常和异常流量背后存在更为复杂的潜在形式，如 NSL-KDD 数据集包含异常流量的 4 种攻击类型：探测攻击(Probe)，拒绝服务(Dos)，用户到根(U2R)和远程到本文(R2L)攻击，并将其分为 40 种更为详细的攻击类型.

同样，根据所使用的服务和协议，正常流量也具有不同的潜在特征. 仅使用潜在变量 $z_1$ 很难准确地表示这些攻击类型的潜在特征. 因此，模型使用潜在变量 $z_2$ 来描述其他信息的详细特征，并假设潜在变量 $z_2$ 服从平均值为 0，标准差为 1 的高斯分布，维数为 50. 因此，模型使用 AE 来降低输入数据的维数并将数据特征保存在两个潜在变量 $z_1$ 和 $z_2$ 中，然后利用顶部鉴别器将分类分布强加给潜在变量 $z_1$ 作为先验分布，用于确保潜在变量 $z_1$ 的分布与分类分布相匹配. 同样，利用底部鉴别器将高斯分布强加给潜在变量 $z_2$ 作为先验分布用于保证潜在变量 $z_2$ 的分布与高斯分布相匹配.

模型在训练时，当有标记的数据可用时，使用标签来训练对抗性自编码器；当使用未标记的数据训练对抗性自编码器时，则假设输入数据由来自分类分布的潜在变量 $z_1$ 和来自高斯分布的潜在变量 $z_2$ 生成. 本文所提模型一旦训练完成，就被用于对新数据进行分类，此时隐藏层中的潜在变量 $z_1$ 表示与输入数据相关联的推断类. 因此，在检测时模型是使用潜在变量 $z_1$ 进行分类的. 半监督对抗性自编码器采用随机梯度下降进行训练，主要分为 3 个阶段：

1) 重构阶段：更新编码器和解码器. 优化AE的参数，使输入和输出数据的重构误差最小. 本文仅在此阶段使用未标记的数据，而AE从该阶段未标记的数据中生成潜在变量 $z_1$ 和 $z_2$.

2) 正则化阶段：训练每个鉴别器来鉴别分类分布的潜在变量 $z_1$ 或样本数据，以及高斯分布的潜在变量 $z_2$ 或样本数据，然后根据鉴别器的检测结果对AE进行优化. 本阶段的训练基于式(4).

3) 半监督分类阶段：对AE进行更新，使标记数据的交叉熵误差最小. 在此阶段，利用标记数据实现半监督学习.

## 3 实验与结果分析

为了评估本文所提入侵检测模型的有效性，在Python 3.6中使用深度学习开源框架Pytorch进行测试，并将测试结果与MSML[13]，STBoost[14]，Bagging-J48[15]等几种半监督学习方法以及LuNet[10]和BAT-MC[16]等深度学习的入侵检测方法进行对比. 所有实验均在Ubuntu 16.04系统，CPU为Intel Pentium@3.50GHz，NVIDIA GeForce GTX1060(6 GB)，RAM 8 GB的环境中进行.

本文所提出的模型使用ADAM优化器. 为了防止过拟合，在训练阶段使用了dropout层和批量归一化. 模型编码器从具有122个特征输入的数据中提取52个重要特征，其中 $z_1$ 为2，$z_2$ 为50. 解码器从具有潜在变量 $z_1$ 和 $z_2$ 的隐藏中间层接收52个输入，然后产生122个输出. 编码器和解码器都具有一个中间的全连接层，其大小为1 000×1 000. 它们在各层之间也使用了dropout层和批量归一化. 分类分布的鉴别器接收两个输入，并产生一个输出(假或真值)，高斯分布的鉴别器接收50个输入也同样产生一个输出(假或真值). 分类分布和高斯分布的鉴别器都具有大小为1 000×1 000的中间全连接层. 它们在各层之间也使用了批量归一化. 所有实验中的批量归一化大小为128，学习率为 $10^{-7}$.

对于对抗AE训练，实验过程中使用NSL-KDD数据集[17]中的训练集作为标记数据，其余样本作为未标记数据. 所有实验使用k-fold($k$ 个折叠，简称 $k$ 折)交叉验证，其中 $k=5$，即将数据集中训练集分为5个部分，训练和验证共进行5次，最后将测试集数据在每个折叠验证精度最高的模型中进行测试，分类结果的平均值为最终测试准确度.

### 3.1 数据集和评估指标

本文选择NSL-KDD数据集作为测试数据，该数据集主要有5个类标签，分为正常类和4种攻击类，其中4种攻击类分别为端口扫描Probe、拒绝服务DoS、远程到本地的攻击R2L以及未经授权的root权限访问U2R，每种攻击类型又划分了多个相应的子类型. 数据集分为41个属性，其中38个数值型属性，3个符号型属性. 本文在利用NSL-KDD数据集进行实验时，采用训练数据集为60%，测试集为40%，具体信息如表1所示.

**表1 实验数据的样本分布**

| 攻击类别 | 训练样本数 | 测试样本数 |
|---|---|---|
| 正常 | 46 310 | 30 880 |
| DoS | 32 057 | 21 328 |
| U2R | 65 | 54 |
| R2L | 2 234 | 1 511 |
| Probe | 8 443 | 5 634 |

为了评估本文所提出方法的有效性，使用准确率 $Acc$，精度 $P$、召回率 $R$、$F1$ 分数和误报率 $FA$ 作为实验过程中的衡量指标. 这些绩效指标可由4种不同的指标计算得到.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = \frac{2 \times R \times P}{R + P} \tag{8}$$

$$FA = \frac{FP}{FP + TN} \tag{9}$$

式(5)－式(9)中：$TP$ 和 $FP$ 分别表示正确预测和错误预测流量为正常类型的样本数，$TN$ 和 $FN$ 分别表示正确预测和错误预测流量为攻击类型的样本数. $TP$，$TN$，$FP$ 和 $FN$ 四者之和为总样本数.

### 3.2　结果分析

1) 为了验证本文所提模型能够在少量标记数据下进行检测的可行性，采用半监督学习的对抗性自编码器与监督学习方法之一的卷积神经网络(CNN)[7]在准确率指标上进行对比. 图 1 给出了标签率(标记数据所占百分比)在 90%，50%，20%，10%，5%和 1%时，本文方法与 CNN 方法的检测准确率结果对比. 从图 1 中可以看出，当数据的标签率在 50%以下时，本文方法的准确率优于 CNN. 根据以上结果，可以认为基于半监督学习的对抗性自编码器入侵检测方法在少量标记数据下进行检测是可行的，因为它可以检测到很高的准确度.

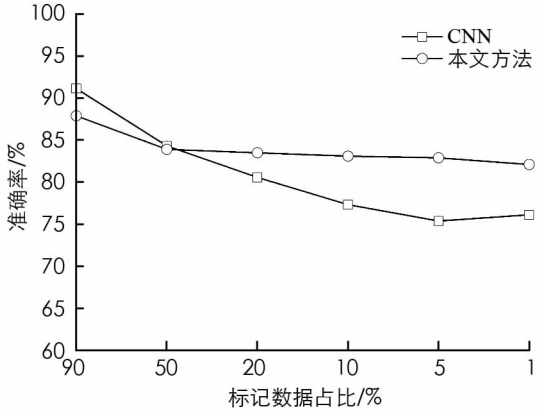


图 1　不同标签率时的结果对比

2) 测试本文所提模型对 NSL-KDD 数据集中 5 种类型的检测结果，如表 2 所示. 从表 2 中可以看出，本文模型能够提供 87.89%的平均准确率. 具体来说，在 NSL-KDD 数据集 5 个类别中，正常和 DoS 的准确率、精度、召回率、$F1$ 分数都很高，误报率很低，但是在 Probe，R2L 和 U2L 这 3 种攻击类别上的检测结果明显低于平均水平，说明本文所提模型需要大量的数据来学习. 当训练数据数量较少时，获得的结果不稳定.

**表 2　所提模型在 NSL-KDD 数据集的测试结果**　%

| 样本类型 | *ACC* | *P* | *R* | *F1* 值 | *FA* |
|---|---|---|---|---|---|
| 正常 | 97.75 | 99.83 | 96.33 | 98.05 | 0.17 |
| DoS | 94.58 | 98.11 | 93.82 | 95.92 | 1.89 |
| U2R | 50.33 | 77.72 | 49.73 | 60.65 | 22.28 |
| R2L | 62.81 | 95.17 | 86.63 | 90.70 | 4.83 |
| Probe | 72.97 | 78.53 | 88.23 | 83.06 | 21.47 |
| 平均值 | 87.89 | 90.06 | 86.76 | 88.34 | 9.94 |

3) 为了进一步验证模型的有效性，图 2 给出了本文所提方法与其他对比方法在准确率和 $F1$ 分数的对比结果. 从图 2 中可以看出，相对于已有的几种半监督学习方法，本文方法在准确率和 $F1$ 分数指标上具有一定的优势，说明将 AE 和 GAN 结合能够实现少量标记数据下的有效检测，但是图 2 中也清楚地显示，本文结果稍逊于基于深度学习的入侵检测方法，该类方法高达 99%的准确率是本文所提模型不能企及的. 而本文提出的半监督机器学习方法的优点在于可以在少量标记数据下实现较高的检测率，不需要太多的标记数据，从而

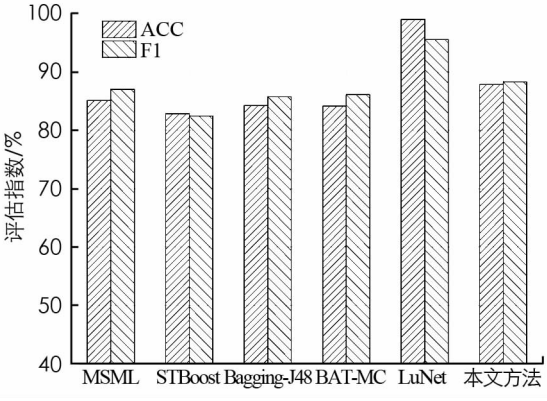


图 2　不同入侵检测方法的测试结果对比

避免了相当数量昂贵的人工任务.

## 4 结 语

本文提出一种基于对抗性自编码器的半监督学习入侵检测算法,用于解决当前无监督学习算法准确度低、误报率高以及有监督学习算法面临的所需训练样本标记困难的问题. 该算法首先利用自编码器对高维输入数据进行特征提取,将重要特征保留在两个潜在变量中,然后服从分类分布和高斯分布的两个潜在变量分别利用上下两个鉴别器进行正则化,最后利用标记数据的交叉熵误差最小来实现分类. 该方法能够利用训练数据集中少量标记数据进行训练,并支持未标记数据. 实验结果表明,本文算法能够在实现高检测准确度、低误报率的同时,减少昂贵的人工任务.

**参考文献:**

[1] BANERJEE J, MAITI S, CHAKRABORTY S, et al. Impact of Machine Learning in Various Network Security Applications [C]//2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). Erode: IEEE, 2019.

[2] WANG Z R, FANG B X. Application of Combined Kernel Function Artificial Intelligence Algorithm in Mobile Communication Network Security Authentication Mechanism [J]. The Journal of Supercomputing, 2019, 75(9): 5946-5964.

[3] KHALAF B A, MOSTAFA S A, MUSTAPHA A, et al. Comprehensive Review of Artificial Intelligence and Statistical Approaches in Distributed Denial of Service Attack and Defense Methods [J]. IEEE Access, 2019, 7: 51691-51713.

[4] KASONGO S M, SUN Y X. A Deep Long Short-Term Memory Based Classifier for Wireless Intrusion Detection System [J]. ICT Express, 2020, 6(2): 98-103.

[5] LI W J, TUG S, MENG W Z, et al. Designing Collaborative Blockchained Signature-Based Intrusion Detection in IoT Environments [J]. Future Generation Computer Systems, 2019, 96: 481-489.

[6] VIEGAS E, SANTIN A, BESSANI A, et al. BigFlow: Real-Time and Reliable Anomaly-Based Intrusion Detection for High-Speed Networks [J]. Future Generation Computer Systems, 2019, 93: 473-485.

[7] POTLURI S, AHMED S, DIEDRICH C. Convolutional Neural Networks for Multi-Class Intrusion Detection System [C]//2018 6th International Conference on Mining Intelligence and Knowledge Exploration(MIKE). Napoca: Springer, 2018.

[8] XU Y Y, LIU Z, LI Y M, et al. Intrusion Detection Based on Fusing Deep Neural Networks and Transfer Learning [M]//Communications in Computer and Information Science. Singapore: Springer, 2020.

[9] WANG W, SHENG Y Q, WANG J L, et al. HAST-IDS: Learning Hierarchical Spatial-Temporal Features Using Deep Neural Networks to Improve Intrusion Detection [J]. IEEE Access, 2018, 6: 1792-1806.

[10] WU P L, GUO H. LuNet: a Deep Neural Network for Network Intrusion Detection [C]//2019 IEEE Symposium Series on Computational Intelligence (SSCI). Xiamen: IEEE, 2019.

[11] ALQATF M, YU L S, AL-HABIB M, et al. Deep Learning Approach Combining Sparse Autoencoder with SVM for Network Intrusion Detection [J]. IEEE Access, 2018, 6: 52843-52856.

[12] MIRZA A H, COSAN S. Computer Network Intrusion Detection Using Sequential LSTM Neural Networks Autoencoders [C]//2018 26th Signal Processing and Communications Applications Conference (SIU). Izmir: IEEE, 2018.

[13] YAO H P, FU D Y, ZHANG P Y, et al. MSML: a Novel Multilevel Semi-Supervised Machine Learning Framework for Intrusion Detection System [J]. IEEE Internet of Things Journal, 2019, 6(2): 1949-1959.

[14] JIANG E P. A Semi-Supervised Learning Model for Intrusion Detection [J]. Intelligent Decision Technologies, 2019, 13(3): 343-353.

[15] PHAM N T, FOO E, SURIADI S, et al. Improving Performance of Intrusion Detection System Using Ensemble Methods and Feature Selection [C]//Proceedings of the Australasian Computer Science Week Multiconference. New York: ACM, 2018.

[16] SU T T, SUN H Z, ZHU J Q, et al. BAT: Deep Learning Methods on Network Intrusion Detection Using NSL-KDD

Dataset [J]. IEEE Access，2020，8：29575-29585.

[17] 曹卫东，许志香，王　静. 基于深度生成模型的半监督入侵检测算法 [J]. 计算机科学，2019，46(3)：197-201.

# Intrusion Detection Algorithm Based on Adversarial Autocoder

BAI Jie-xian[1]，　JU Lei-ming[2]

1. *Computer Information Engineering College，Shanxi Technology and Business College，Taiyuan 030006，China；*

2. *Software School，Nanyang Institute of Technology，Nanyang Henan 473000，China*

**Abstract**：Aiming at the problems of low accuracy and high false alarm rate for intrusion detection algorithm based on unsupervised learning，and high cost of training samples required by supervised algorithm，an intrusion detection algorithm based on adversarial autocoder has been proposed. This is a semi-supervised learning algorithm，which only needs a small amount of labeled data in the training data set for training，and supports unlabeled data in the training data set，so as to improve the performance. Firstly，the autocoder reduces the dimensionality of the input data by extracting important features as latent variables；secondly，it uses the generative adversarial network to make the latent variables of the autocoder follow an arbitrary distribution for regularization；and finally，it uses the cross entropy loss of labeled data to achieve the classification of semi-supervised learning. Experimental results show that，compared with other algorithms，the proposed algorithm has certain advantages in detecting a limited number of labeled samples，which can achieve high accuracy and low false alarm rate，while reducing the demand for labeled data.

**Key words**：intrusion detection algorithm；autocoder；generative adversarial network；semi-supervised learning

**责任编辑**　夏　娟