计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200600022

# 基于变分自编码器的不平衡样本异常流量检测

张仁杰　陈　伟　杭梦鑫　吴礼发
南京邮电大学计算机学院、软件学院、网络空间安全学院　南京 210023
(zrj9582346@163.com)

**摘　要**　随着机器学习技术的快速发展,越来越多的机器学习算法被用于攻击流量的检测与分析,然而攻击流量往往只占网络流量中极小的一部分,在训练机器学习模型时存在训练集正负样本不平衡的问题,从而影响模型训练效果。针对不平衡样本问题,文中提出了一种基于变分自编码器的不平衡样本生成方法,其核心思想是在对少数样本进行扩充时,不是对全部进行扩充,而是分析这些少数样本,对其中最容易对机器学习产生混淆效果的少数边界样本进行扩充。首先,利用KNN算法筛选出少数类样本中与多数类样本最近的样本;其次,使用DBSCAN算法对KNN算法筛选出的部分样本进行聚类处理,生成一个或多个子簇;然后,设计变分自编码网络模型,对DBSCAN算法区分出的一个或多个子簇中的少数类样本进行学习扩充,并将扩充后的样本加入原有样本中用于构建新的训练集;最后,利用新构建的训练集来训练决策树分类器,从而实现异常流量的检测。选择召回率和$F1$分数作为评价指标,分别以原始样本、SMOTE生成样本、SMOTE改进方法生成样本和文中所提方法生成样本为训练集进行对比实验。实验结果表明,在4种异常类型中,采用所提算法构造训练集训练的决策树分类器在召回率和$F1$分数上都有提升,$F1$分数相比原始样本及SMOTE方法最高提升了20.9%。

**关键词**:异常流量;过采样;变分自编码器;不平衡样本;KNN;DBSCAN
**中图法分类号**　TP391

## Detection of Abnormal Flow of Imbalanced Samples Based on Variational Autoencoder

ZHANG Ren-jie,CHEN Wei,HANG Meng-xin and WU Li-fa
School of Computer Science,Nanjing University of Posts and Telecommunications,Nanjing 210023,China

**Abstract**　With the rapid development of machine learning technology,more and more machine learning algorithms are used to detect and analyze attack traffic. However,attack traffic often accounts for a very small portion of network traffic. When training machine learning models,there is often a problem of imbalance between the positive and negative samples of the training set,which affects model training effect. Aiming at the problem of imbalanced samples,an imbalanced sample generation method based on variational auto-encoder is proposed. The idea is that when expanding imbalanced samples,not all of them are expanded. But imbalanced samples are analyzed,and a small number of boundary samples that are most likely to have confusion effects on machine learning are expanded. First,the KNN algorithm is used to screen the samples that are closest to the majority of samples;second,DBSCAN algorithm is used to cluster the partial samples selected by the KNN algorithm to generate one or more sub-clusters;then,a VAE network model is designed to learn and expand the few samples in one or more sub-clusters distinguished by the DBSCAN algorithm. The expanded samples are added to the original samples to build a new training set;finally,the newly constructed training set is used to train decision tree classifier to detect abnormal traffic. The recall rate and $F1$ score are selected as the evaluation indicators. The original sample,the SMOTE-generated sample and our sample are compared. The experimental results show that the decision tree classifier trained using the proposed method in this paper has improved the recall rate and $F1$ score among the four types of anomalies. The $F1$ score is up to 20.9%,which is higher than the original sample and the SMOTE method.

**Keywords**　Abnormal flow,Oversampling,Variational auto-encoder,Imbalanced sample,KNN,DBSCAN

到稿日期:2020-06-02　返修日期:2020-08-23　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划(2019YFB2101704)
This work was supported by the National Key Research and Development Project(2019YFB2101704).
通信作者:陈伟(chenwei@njupt.edu.cn)

## 1 引言

随着互联网的普及，网络流量呈现爆炸性的增长趋势，2019 年 9 月，中国互联网信息中心发布《中国互联网络发展状况统计报告》，报告指出，截至 2019 年 6 月，我国互联网人口规模已达 8.47 亿[1]。随着网络高速发展，网络安全问题日趋严峻，近年来网络安全问题的种类不断增多，影响的用户范围也越来越广[2]。仅在 2017 年上半年，全球物联网攻击事件就增长了 280%[3]。2017 年 9 月，某物联网安全研究公司在蓝牙协议中发现了 8 个 0day 漏洞，该漏洞波及全球数十亿台蓝牙设备，同年 10 月，Wi-Fi 网络中广泛使用的 WAP2 安全协议被曝出了重大安全漏洞，令几乎所有的 Wi-Fi 设备都遭受了影响[4]。随着机器学习技术的发展，其在网络流量异常检测领域中逐渐发挥了重要作用[5]，但攻击流量通常只占网络流量的极小一部分，而用于机器学习训练的异常流量样本往往存在样本数据不平衡的问题[6]。

不平衡样本数据指样本数据类别间分布明显不均衡的样本，其中样本数量较多的类别为多数类，样本数量较少的类别为少数类[7]。不平衡样本普遍存在于多个领域中，如工业领域中的故障检测[8]、图片识别、信用卡欺诈检测等[9]。由于网络流量中异常流量样本较少而正常流量样本较多，因此采用多数类样本时机器学习方法训练出的分类器的准确率偏高，少数类样本的检测准确度很低。然而，少数类样本也就是样本中的异常流量样本才是异常流量检测关注的重点，且异常流量的漏报危害总是大于误报的。

本文提出了一种利用变分自编码器进行少数类样本扩充，其核心思想是在对少数样本进行扩充时，不是对全部进行扩充，而是分析少数样本，对其中最容易对机器学习产生混淆效果的少数边界样本进行扩充，最终实现网络流量异常检测的方法。首先，基于 KNN 设计了一种少量样本选择算法。KNN 算法是一种便于实现的分类算法，利用基于 KNN 的少量样本选择算法找出少数类样本中与多数类样本最近的样本，由于这些样本的空间距离与正常样本的空间距离接近，容易在模型训练中与正常样本混淆，因此这些样本是需要扩充的关键部分。其次，使用基于 DBSCAN 设计的少量样本类内聚类算法。DBSCAN 是一种基于密度的聚类算法，它的作用是将样本划分为多个簇，每个簇中都有相同类别的样本，即使数据中存在噪声也可以发现任意形状的聚类。利用基于 DBSCAN 算法设计的类内聚类算法对 KNN 少量样本选择算法筛选出的部分样本进行聚类处理，找出少量样本间的差异，从而生成一个或多个子簇，提高生成模型输入样本的质量。最后，设计变分自编码网络模型，对 DBSCAN 类内聚类算法区分出的一个或多个子簇中的少类样本进行学习扩充，并将扩充后的样本加入原有样本中构建新的训练集。利用扩充后的训练集训练机器学习决策树分类器，从而实现异常流量的检测。使用原始未扩充训练样本、SMOTE 扩充训练样本与本文算法扩充的训练样本分别训练决策树分类器，经过对比实验，验证了本文方法的有效性和可行性。

## 2 相关研究

为了提高对不平衡样本数据分类的能力，近年来学术界进行了大量的研究，总结来看其主要分为数据层面和算法层面。在数据层面，对不平衡样本进行直接处理，包括过采样和下采样两种方法。过采样方法是对少数类样本进行扩充处理，主要有随机过采样和 SMOTE[10] 两种方法。随机过采样的主要方法是对少数类样本进行随机复制，使少数类样本的数量接近或者达到多数类样本的数量。随机过采样方法较为简单也便于实现，但是这种方法增大了机器学习模型对于训练样本过拟合的可能性，往往无法产生好的效果。SMOTE 方法是近年来 Chawla 等提出的一种少数类样本扩充方法，相比随机过采样方法中简单地对少数类样本进行复制，SMOTE 在两个少数类样本间通过线性插值的方法合成新的样本，较为有效地解决了过拟合的问题。由 SMOTE 改进产生了诸多算法，如 ADASYN[11]，Borderline-SMOTE[12]，SMOM[13]，G-SMOTE[14] 等。下采样方法是对多数类样本进行随机删除处理，使其样本数量减少到或接近于少数类样本，但是该方法可能造成大量有用信息丢失，容易导致分类器能力下降。在算法层面，主要通过引入损失函数和提升分类器的健壮性来提高对不平衡样本的分类能力，主要算法有代价敏感学习[15]、集成学习[16] 和提升算法[17]。

近年来，在深度学习领域，深度学习生成模型取得了重大进展[18]，变分自编码器（Variational Autoencoder，VAE）[19] 便是其中的一种。变分自编码器作为一种特殊的自编码器模型，一经提出便受到了极大欢迎，被广泛应用于图像生成领域。本文参考变分自编码器在图像生成领域中的应用方法，利用变分自编码器重构原始输入样本概率密度分布的能力，对网络异常检测领域中的少量样本进行扩充，以提高异常检测效果。

## 3 变分自编码器的原理

自编码器（autoencoder）是一种深度学习模型，是一种在半监督和非监督学习中使用的神经网络，其功能是对输入信息进行表征学习，目前主要应用于特征降维[20]、图像分类降噪[21] 等领域。自编码器是一种学习目标与输入相同的神经网络模型，其基本结构有 3 层，即输入层、隐藏层（中间状态）和输出层。自编码器由编码器和解码器两部分组成[22]，其中输入层到隐藏层（中间状态）的部分为编码器，隐藏层（中间状态）到输出层的部分为解码器。如图 1 所示，编码器将输入数据 $X$ 编码生成集合 $X'$，即 $X'=e(x)$；解码器将编码器生成的 $X'$ 解码还原成 $X$，即 $X=d(x')=d(e(x))$。

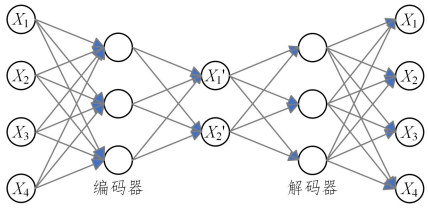



图 1 自编码器结构示意图

Fig. 1 Schematic diagram of AE

变分自编码器（VAE）是一种由 Kingma 等于 2014 年提出的深度学习生成模型，它的名称与自编码器接近，网络结构也十分相似，目前主要应用于推荐系统[23]、网络入侵检

测[24]、人脸识别[25]等领域。与传统自编码器不同，变分自编码器学习的不再是样本的个体，而是通过建立样本的概率密度分布模型来学习样本的分布规律。类似于自编码器的模型结构，变分自编码器模型同样由编码器和解码器两部分组成。在 VAE 中，编码器用于建立原始输入数据的变分推断，生成隐变量的变分概率分布，因此编码器通常也称为推断网络；解码器根据推断网络生成的隐变量变分概率分布，生成与原始数据接近的概率密度分布，因此解码器通常也称为生成网络[26]。如图 2 所示，推断网络将原始数据编码生成隐变量 $Z$，生成网络根据隐变量 $Z$ 还原生成原始数据的近似概率分布。

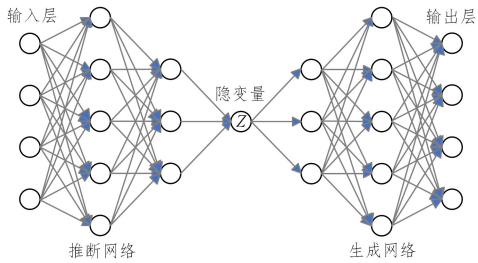



图 2 变分自编码器结构示意图

Fig. 2 Schematic diagram of VAE

由于隐变量 $Z$ 的分布不可直接观测，无法直接使用最大期望算法进行变分推断求解，因此变分自编码器模型在推断网络中引入了一个识别模型 $p_{\emptyset}(z|x)$，来代替无法确定的真实概率分布 $q_{\theta}(z|x)$，这样模型 $p_{\emptyset}(z|x)$ 就可以作为变分自编码器的推断网络部分，条件分布 $q_{\theta}(x'|z)$ 作为生成网络部分。为了使识别模型与真实的概率密度分布近似相等，变分自编码器使用 KL 散度计算二者的差异大小，并通过优化参数 $\emptyset$，$\theta$，使 KL 散度值。最小化

因此，变分自编码模型有两个学习目标：

(1)最小化样本的重构损失。

(2)最小化 KL 散度值。

模型的 KL 散度计算式如式(1)所示：

$$KL(p(x)|q(x))=\int p(x)\frac{p(x)}{q(x)}\mathrm{d}x=E_{x\sim p(x)}\log\frac{p(x)}{q(x)} \tag{1}$$

其中，$p(x)$和 $q(x)$表示两个概率分布，KL 散度用于计算识别模型和生成的概率分布间的差异性。

## 4 基于变分自编码器的异常流量检测方法

### 4.1 KNN 少量样本选择算法

少量样本生成算法对不平衡样本数据集中的少量样本进行生成时，并非需要以全部少量样本为整体进行样本扩充。本文选择的机器学习方法为决策树模型，该模型利用样本中不同特征的数值区别，提取出一系列规则，按照提取规则进行样本类别划分。少量样本中的部分样本与其他类别的样本在特征数值上存在明显差异，无需进行样本扩充，决策树模型不仅能很好地识别出该类样本的特点，也能够在模型训练完成后的检测过程中准确地发现这类少量样本。因此，这些样本不会对最终决策树分类器的判别效果产生较大影响。然而，还有部分少量样本在空间距离(欧氏距离)上远离同类样本而与其他多数类样本更加接近，由于这类样本的数量较少，决策树模型在训练时无法充分学习这类样本的特征，从而导致模型泛化能力差，将直接影响分类器的判别效果。因此，本文首先设计了 KNN 少量样本选择算法，在少量样本中选择部分在空间距离上与多数类样本更加接近的一部分样本，提取这一部分样本进行扩充。

相比直接选择全部少量样本为整体的样本扩充方法，经过 KNN 少量样本选择算法筛选的样本数量更少，但利用筛选的样本对变分自编码器进行训练，可以让变分自编码器更加准确地学习到需要扩充的少量样本的概率密度分布，对提升决策树分类器的效果有更大的帮助。本文筛选这部分少量样本的规则为：若某个少量样本的 $K$ 个近邻中超过一半的样本为多数类样本，则将这个少量样本取出来并定义为需要扩充的少量样本。其中，$K$ 并非是固定的，根据对比实验在不同类样本中选择不同的 $K$ 值，可实现检测效果提升的最大化。该算法的具体步骤如算法 1 所示。

**算法 1** KNN 少量样本选择算法

输入：正常样本 $S_n$，某一类少量异常样本 $S_a$，与选择点距离最小的样本数量 K

输出：少量样本中与正常样本更加接近的样本 D(DBSCAN 类内聚类算法输入数据)

```
S=Sn∪Sa
Dis=0
D=0
num_att,num_nor=0
for each sample p in Sa
    calculate distance to S
    add distance to Dis
    sort Dis from small to large
    for i= 1 to K
      if Dis[i] in Sa then
        num_att=num_att+1
      else
        num_nor=num_nor+1
      end if
    end for
    if num_nor >=num_att then
        add p to D
    end if
end for
return D
End
```

### 4.2 DBSCAN 类内聚类算法

在训练分类器模型时往往只考虑了不同类别样本之间的差异，而忽略了同类样本内部的差异。虽然经过 KNN 少量样本选择算法筛选后的样本数量很少，但是这些样本仍然存在较大的差异性。即使是同一种类型的异常也会有不同的异常特征，如果将未分类的所有样本送入变分自编码器中用于训练生成样本的概率密度分布，那么利用这一概率密度分布模型生成的新样本将无法很好地拟合原始样本，更多的合成样本将分布在原始样本的边缘。在过采样的过程中，生成样

本的质量越高，训练的模型检测效果就会越好，如果生成样本的质量很低，大量生成样本分布在原始样本的边缘，这样不仅无法提升分类器对少量样本的检测效果，反而会影响对多数类样本分类的准确度。因此，在将原始样本送入变分自编码器模型进行少量样本生成前，需要先对原始样本内部进行分类处理。

本文采用 DBSCAN 聚类算法进行少量样本的类内聚类，该聚类算法是一种基于密度的聚类算法。与划分聚类和层次聚类算法不同，它的作用是将样本划分为多个簇，每个簇中都有相同类别的样本，即使数据中存在噪声也可以发现任意形状的聚类。与传统的 K-Means 方法相比，DBSCAN 可以发现任意形状的簇类，无需事先设定要形成簇类的数量，这也是本文选取 DBSCAN 作为聚类算法的主要原因。使用 DBSCAN 将经过 KNN 少量样本选择方法筛选后的样本聚类，调整 DBSCAN 中 *Eps*，*Mins* 等参数，使得样本产生一个或多个聚类，每个聚类为一个子簇。最后将每个子簇分别取出，使用变分自编码器对少量样本的每个子簇分别进行训练，同时生成新的样本。该算法的具体步骤如算法 2 所示。

**算法 2** DBSCAN 类内聚类算法

输入：KNN 少量样本选择算法筛选后的部分样本 D，同一聚类中两个样本的最大距离 Eps，同一聚类集合中的最小样本数 Mins

输出：基于密度的聚类 S(变分自编码器的输入数据)

```
S=0;
for each unvisited point p in D
   mark p as visited
   N=get Neighbours (p, Eps)
   if sizeof (N) <Mins then
      mark p as Noise
   else
      S=next cluster
      ExpandCluster(p,N,S,Eps,Mins)
   end if
end for
ExpandCluster(p,N,S,Eps,Mins)
add p to cluster S
for each unvisited point p′ in N
   mark p′ as visited;
   N′=getNeighbours (p′,Eps)
   if sizeof(N′)≥Mins then
      N=N+N′
   end if
   if p′ is not member of any cluster
      add p′ to cluster S
   end if
end for
return S
End
```

### 4.3 变分自编码器少量样本生成算法

不平衡样本数据集经过上述算法的处理后，可以得到少量样本中需要扩充的部分样本。本文研究使用 UNSW-NB15 数据集，该数据集包含 42 个特征和 1 个类别标签，为方便处理，本文使用其中 39 个数值特征作为变分自编码器的训练特征。神经网络在层数选择上需要综合考虑输入特征维度及训练时间，一般情况下层数越多，训练效果就越好，但训练时间加长，同时过多的层数也可能带来模型过拟合问题。由于本文选择的特征只有 39 个，特征数量较少，在输入样本特征数量较少的情况下模型无需过多的隐藏层数量。因此，本文设计了一个包含隐变量层在内的 4 层全连接神经网络，其中第 1－2 层为变分自编码器的编码器，也就是推断网络部分；第 3－4 层为变分自编码器的解码器，也就是生成网络部分。在隐藏层节点的选择方面，先前的大量文献多采用式(2)来计算隐藏层节点的数量。其中，$N_x$ 为输入层节点数，$N_y$ 为输出层节点数，$N_z$ 为计算得到的隐藏层节点数量。

$$N_z=\sqrt{N_xN_y} \tag{2}$$

当本文输入层节点数为 39、输出层节点数为 2 时，通过计算可知，隐藏层的节点数应为 9。因此，本文设计的变分自编码器网络模型的第 1 个全连接层输入层为 39 个维度，输出层为 9 个维度；第 2 个全连接层并列连接了两个输出网络，分别输出输入样本的均值与方差，每个网络都输出两个维度的输出，然后将两个维度输出计算为一个 2 维度的数据，计算式如式(3)所示：

$$Z=mean+N(0,1)\sqrt{e^{sigma}} \tag{3}$$

其中，*mean* 为均值，*sigma* 为方差，$N(0,1)$为符合标准正态分布的一个数。

接着，将上述数据输入一个以 2 维度为开始的解码部分，解码器即生成网络部分包含两个全连接层的解码器。第 1 层输入层为 2 个维度，输出层为 9 个维度；第 2 层输入层为 9 个维度，输出层为 39 个维度。详细的网络层次结构如图 3 所示。

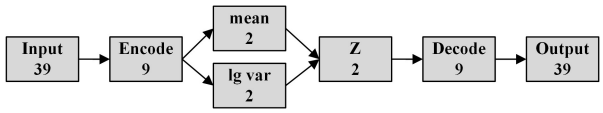



图 3 本文变分自编码器的网络结构

Fig. 3 VAE network structure in this paper

在生成样本前首先对变分自编码器模型进行训练，少量样本经过 KNN 少量样本选择算法和 DBSCAN 类内聚类算法的处理后，得到需要扩充的部分样本，将这些样本送入变分自编码器的输入层部分。由于本文实验采用的 UNSW-NB15 数据集中的少量样本数量较少，经过上述两种算法处理后，训练样本的数量均不足 1 000，因此本文设计的变分自编码器模型在训练时将 *batchsize* 大小设置为样本整体大小。较大的 *batchsize* 有助于提高内存利用率，加快数据处理的速度。同时，合理增大 *batchsize* 的取值可以使模型训练时梯度下降的方向更加准确，防止模型最终陷入不同的局部最优值，难以收敛。对变分自编码器模型迭代训练多次，直到损失函数值稳定，则停止迭代训练，保存该模型。

变分自编码器模型训练完成后，调用保存的生成网络部分进行样本生成，生成网络的输入部分需要输入一个 2 维度的数据。本文设计的输入数据中，维度 1 的输入数据是将整个高斯分布数据集从小到大排列，取出其中 0.1 至 0.9 的区间，生成 $m$ 个随机数。维度 2 的输入数据同样是将整个高斯

分布数据集从小到大排列，取出其中 0.1 至 0.9 的区间，生成 $n$ 个随机数。这样最终产生了 $m*n$ 个符合标准高斯分布且分布区间为 0.1 至 0.9 的 2 维数据。将这些数据分别输入并保存到生成网络模型中，循环输入 $m*n$ 次后，即可得到 $m*n$ 个按照原始训练样本的概率密度分布的新样本。将新生成的样本与原始训练样本合并，即可得到经过变分自编码模型扩充后的新的训练样本集。

## 5 实验与分析

### 5.1 数据集

为了验证算法的有效性，本文选择了 UNSW-NB15 入侵检测数据集作为实验数据集[27]。该数据集由澳大利亚网络安全实验室利用 IXIA PerfectStrom 工具生成，比传统的 KDD-99 和 NSL-KDD 入侵检测数据集更能反映现代网络流量的特点。该数据集共有 42 个特征，详细的特征描述如表 1 所列，本文选择其中 39 种数值特征作为训练使用的特征。数据集中包含一个类别标签，标签区分了 9 种不同的异常类型和 1 个正常类型，包括 Analysis，Backdoor，DoS，Exploits，Fuzzers，Generic，Reconnaissance，Shellcode，Worms，Normal。详细的异常类型的含义如表 2 所列。

表 1 UNSW-NB15 数据集的特征描述

Table 1 Characterization of UNSW-NB15 dataset

| 特征类别 | 特征名称 |
|---|---|
| 基础特征 | state，dur，sbytes，dbytes，sttl，dttl，sloss，dloss，service，sload，dload，spkts，dpkts |
| 连接特征 | swin，dwin，stcpb，dtcpb，smeansz，dmeansz，trans_depth，res_bdy_len |
| 时间特征 | sjit，djit，sintpkt，dintpkt，tcprtt，synack，ackdat |
| 额外的生成特征（通用特征） | is_sm_ips_ports，ct_state_ttl，ct_flw_http_mthd，is_ftp_login，ct_ftp_cmd |
| 额外的生成特征（连接特征） | ct_srv_src，ct_srv_dst，ct_dst_ltm，ct_src_ltm，ct_src_dport_ltm，ct_dst_sport_ltm，ct_dst_src_ltm |

表 2 UNSW-NB15 数据集的异常类型描述

Table 2 UNSW-NB15 data set exception type description

| 异常类型 | 描述 |
|---|---|
| Analysis | 通过端口、web、邮件方式入侵 |
| Backdoor | 通过绕过系统安全性控制而获取系统访问权限 |
| DoS | 通过直接或间接耗尽被攻击者资源使其无法提供服务或资源访问 |
| Exploits | 通过触发一个或多个漏洞进而控制目标系统 |
| Fuzzers | 通过在系统安全漏洞中输入大量随机数据使系统崩溃 |
| Generic | 不考虑分组密码配置而使用哈希函数对每个分组密码进行碰撞 |
| Reconnaissance | 通过逃避安全控制收集计算机信息 |
| Shellcode | 通过发送利用特定漏洞的代码控制目标机器 |
| Worms | 通过网络传播主动攻击计算机的恶性病毒 |

### 5.2 数据集预处理

UNSW_NB15 数据集中一共包含了 2540044 条记录，本文对其中的 9 种异常类别与 1 种正常类别按相同比例，从中抽取了 257673 条数据，并按 70%与 30%的比例将样本划分为训练集和测试集。在构造完成后的训练数据集中，各种类别异常占训练集整体的比值如表 3 所列。可以看出，Analysis，Backdoor，Shellcode，Worms 这 4 种类型异常在所有训练样本中所占的比例均不足 1%，符合不平衡样本中少量样本的特点。因此，本文选择这 4 种类型的异常样本作为少量样本，测试本文提出的少量样本生成算法在提升异常检测效果方面的有效性。

表 3 训练集各种类型异常数量及其所占比例

Table 3 Number of abnormal types and their proportions on training set

| 异常类型 | 数量 | 占比 |
|---|---|---|
| Analysis | 1874 | 1% |
| Backdoor | 1631 | 0.9% |
| DoS | 11448 | 6.3% |
| Exploits | 31668 | 17.2% |
| Fuzzers | 16973 | 9.4% |
| Generic | 41210 | 22.8% |
| Reconnaissance | 9791 | 5.4% |
| Shellcode | 1058 | 0.5% |
| Worms | 122 | 小于 0.1% |

### 5.3 性能评估指标

针对不平衡样本数据的特点，不能简单地选择准确率作为单一评价指标。由于少量样本类别中的训练样本过少，机器学习模型在训练中只能去拟合这一部分的异常特征，导致在测试时准确率很高但召回率很低。召回率和 $F1$ 分数这两项指标往往能更准确地反映出少量样本的扩充效果，因此本文选择召回率和 $F1$ 分数（$F1$ score）作为评价指标。$F1$ 分数是一种在统计学中用于衡量二分类模型准确率的一种指标，近年来被广泛应用于机器学习模型的效果评估。$F1$ 分数同时兼顾了分类模型的准确率和召回率，可以看作是二者的一种调和平均，它的最大值是 1，最小值是 0。利用混淆矩阵表示不平衡数据的分类结果，如表 4 所列。

表 4 混淆矩阵

Table 4 Confusion matrix

| 分类 | 实际异常 | 实际正常 |
|---|---|---|
| 预测异常 | TP | FP |
| 预测正常 | FN | TN |

根据表 4，准确率（PR）、召回率（RC）、$F1$ 分数（FS）的定义如式（4）—式（6）所示：

$$PR=\frac{TP}{TP+FP} \tag{4}$$

$$RC=\frac{TP}{TP+FN} \tag{5}$$

$$FS=2\frac{PR*RC}{PR+RC} \tag{6}$$

### 5.4 KNN 参数选择

在 KNN 少量样本选择算法中，$K$ 值的选择对于最终检测效果的提升有着直接的影响，因为 $K$ 值的不同使得 KNN 算法在少量样本中筛选出的样本分布和数量不同，从而导致变分自编码器输入的样本不同，产生不同的概率密度分布。为了确定最佳的 $K$ 值，本文设计了一种实验方法，以寻找出最佳的 $K$ 参数。首先，限定 $K$ 的取值范围为 3～9，利用相同的 KNN 算法和相同的 $K$ 值分别对训练集和测试集中的同类样本进行筛选。在训练集与测试集中滤除经过筛选的（即需要进行扩充的）少量样本，利用剩余样本构建新的训练集与测试集，并利用新构造的训练集训练决策树模型，使用新的测试集对模型的准确度进行测试。新产生的训练集和测试集仅用

于该实验，在后续实验中评估模型效果时仍使用原始训练集和原始测试集作为训练样本和测试样本。经过了 KNN 样本筛选，对留下的样本进行测试，如果准确度更高就代表这一部分的少量样本预测精确度已经很高，无需再使用变分自编码器对这一部分样本进行扩充处理。经过实验，各 $K$ 值下几种类型的少量样本检测结果的 $F1$ 分数如表 5 所列，我们选择其中分数最高的 $K$ 值作为该类别样本 KNN 的 $K$ 参数。

表 5 不同 $K$ 值对应的各类异常 $F1$ 分数

Table 5 Various abnormal $F1$ scores corresponding to different $K$ values

| 异常类型 | $K=3$ | $K=4$ | $K=5$ | $K=6$ | $K=7$ | $K=8$ | $K=9$ |
|---|---|---|---|---|---|---|---|
| Analysis | 0.9563 | 0.9759 | 0.9684 | 0.9893 | 0.9707 | 0.9860 | 0.9795 |
| Backdoor | 0.9665 | 0.9873 | 0.9825 | 0.9813 | 0.9869 | 0.9905 | 0.9885 |
| Shellcode | 0.8342 | 0.8636 | 0.8722 | 0.8837 | 0.8741 | 0.8532 | 0.8866 |
| Worms | 0.7606 | 0.8148 | 0.8286 | 0.8525 | 0.8947 | 0.8519 | 0.9014 |

### 5.5 DBSCAN 参数选择

相比 K-Means 算法，DBSCAN 聚类算法不需要事先设定要形成簇类的数量，就可以发现任意形状的簇类。但 DBSCAN 仍然有两个核心参数需要设置，即 *eps* 和 *mins*。*eps* 指同一聚类集合中两个样本的最大距离，*eps* 越小，同一簇类中的样本相关性就越高，反之则相关性越低。为了防止同一类样本生成的簇类过多，同时保证同一簇类样本有足够的相似性，本文将各种异常类型的 *eps* 值均设置为 0.5～2.0。*mins* 指同一聚类集合中最小样本的数量，*mins* 越小，最小簇类中的样本数量越少，样本间的相关性就越高，反之最小簇类样本数量越多，相关性就越低。与 *eps* 值的设置相似，为了保证最终生成样本的效果，本文在多次实验后将各种异常类型的 *mins* 值设置为 5～10。各类异常样本详细的 *eps*，*mins* 参数如表 6 所列。

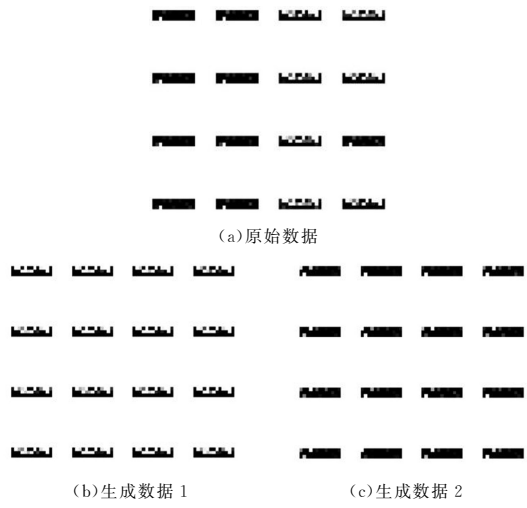
表 6 各类异常 *eps*，*mins* 参数

Table 6 Various abnormal *eps*，*mins* parameters

| 异常类型 | *eps* | *mins* |
|---|---|---|
| Analysis | 0.5 | 6 |
| Backdoor | 0.5 | 7 |
| Shellcode | 1 | 6 |
| Worms | 1.7 | 5 |

### 5.6 本文方法与其他方法对比

为了验证本文方法进行少量样本数据扩充的有效性，先将部分原始样本与生成样本转换为灰度点图的形式，将生成样本与原始样本的相似程度进行可视化。由于本文选取样本中的 39 个特征，因此选择将样本生成为 3 行 13 列的灰度点图，通过灰度直观展示了样本中每个特征的数值大小。本文使用 python 中的 matplotlib 包生成灰度点图，实验以 Shellcode 恶意类型为例，使用 KNN 少量样本选择算法，首先在 Shellcode 类恶意样本中筛选出需要生成的部分样本，随机在筛选出的部分样本中选择 9 条恶意样本生成灰度点图；接着使用 DBSCAN 类内聚类算法对需要生成的样本进行聚类，聚类后得到两个子簇，分别对两个子簇内的样本使用本文设计的变分自编码器模型进行样本生成；最后，分别在扩充后产生的两类样本中，同时随机选择 9 条样本生成灰度点图，两者生成的灰度点图对比如图 4 所示。



(a)原始数据　(b)生成数据 1　(c)生成数据 2

图 4 原始样本与生成样本灰度点图对比

Fig. 4 Gray point map comparison of original sample and generated sample

图 4(a)给出了 Shellcode 中选取的 9 条原始样本的灰度点图形态，图 4(b)和图 4(c)给出了采用本文算法生成的 9 条 Shellcode 类恶意样本。从图 4(a)中可以看出，原始样本大致有两种灰度点图样式，意味着在 Shellcode 类恶意样本中存在两种不同的恶意特征。通过对比观察可以发现，图 4(b)和图 4(c)采用本文算法生成的 Shellcode 类恶意样本的形态与原始样本的形态较为接近，生成的恶意样本灰度点图同样反映出了原始数据中的两种灰度点图样式，可以看出通过变分自编码器生成的恶意样本较为符合原始样本的特征。

接下来，对比本文提出的少量样本生成算法的生成样本与 SMOTE 算法及其改进算法(Borderline-SMOTE)的生成样本在机器学习模型中的异常检测效果。本文使用 python 语言中的 Sklearn 包搭建决策树分类模型，实验均使用其中的 CART 决策树算法。分别使用 Analysis＋Normal，Backdoor＋Normal，Shellcode＋Normal，Worms＋Normal 4 种恶意类型训练集训练决策树模型，以构建二分类的决策树分类器。使用相同的测试集对原始样本、SMOTE 扩容样本、Borderline-SMOTE 扩容样本和本文算法扩充样本训练出的决策树分类器进行测试，记录其中的召回率和 $F1$ 分数，以此判断不同训练集对决策树分类器检测效果的影响。最终的实验结果如表 7、表 8 所列。

表 7 不同方法的召回率对比

Table 7 Comparison of recall methods by different methods

| 异常类型 | 原始样本 | SMOTE | B-SMOTE | 本文算法 |
|---|---|---|---|---|
| Analysis | 0.8804 | 0.8917 | 0.8754 | 0.8966 |
| Backdoor | 0.9441 | 0.9556 | 0.9584 | 0.9599 |
| Shellcode | 0.8102 | 0.8653 | 0.7807 | 0.8319 |
| Worms | 0.7692 | 0.8077 | 0.7503 | 0.8653 |

表 8 不同方法的 $F1$ 分数对比

Table 8 Comparison of $F1$ scores of different methods

| 异常类型 | 原始样本 | SMOTE | B-SMOTE | 本文算法 |
|---|---|---|---|---|
| Analysis | 0.8826 | 0.8801 | 0.8776 | 0.8978 |
| Backdoor | 0.9435 | 0.9468 | 0.9502 | 0.9613 |
| Shellcode | 0.7884 | 0.8033 | 0.7662 | 0.8366 |
| Worms | 0.7767 | 0.7434 | 0.7090 | 0.8571 |

由表 7、表 8 给出的实验结果可以看出，本文提出的样本扩充方法确实提升了决策树分类器的检测效果。在 4 种异常类型中，采用本文算法生成数据训练的决策树分类器相比原始数据训练的分类器，在召回率和 $F1$ 分数上都有提升。其中，Worms 类型的异常检测效果提升最为明显，召回率提高了约 12.5%，$F1$ 分数提升了约 10.4%。对比其他传统生成算法，本文算法在 4 种恶意类型的检测中召回率、$F1$ 分数大多高于对比实验中的 SMOTE 和 Borderline-SMOTE 方法，仅有 Shellcode 类型异常中，本文算法的召回率低于 SMOTE 算法。其中，在 Shellcode 和 Worms 两类异常中本文算法相比 SMOTE、Borderline-SMOTE 方法的 $F1$ 分数提升更为明显，分别提升了约 9.2% 和 20.9%。观察实验结果还可以发现，在 Analysis 和 Worms 两类异常中，虽然使用 SMOTE 方法生成数据训练的分类器的召回率相比原始数据训练的分类器有所提升，但 $F1$ 分数却不及原始数据训练的分类器，Borderline-SMOTE 方法在 Shellcode 和 Worms 类异常样本中甚至出现了召回率、$F1$ 分数均低于原始数据训练的分类器。可见，使用 SMOTE 和 Borderline-SMOTE 方法对这两种少量样本进行处理可能会使得分类器的检测能力下降，而使用本文算法生成数据训练的决策树分类器则不存在这样的问题。

最后，对比本文设计的少量样本扩充方法、原始方法、SMOTE 方法和 SMOTE 改进方法的时间消耗。4 种方法均采用相同的随机森林模型作为预测模型，分别对比不同方法全流程所消耗的时间，其中本文方法的时间包括 KNN 少量样本选择算法、DBSCAN 类内聚类算法、变分自编码器少量样本生成算法及最终的随机森林预测算法所消耗的总时间，最终的对比结果如表 9 所列。

表 9 不同方法的时间消耗对比

Table 9 Comparison of time consumption of different methods

(单位：s)

| 异常类型 | 原始样本 | SMOTE | B-SMOTE | 本文算法 |
|---|---|---|---|---|
| Analysis | 7.69 | 10.07 | 14.33 | 302.12 |
| Backdoor | 8.24 | 10.18 | 13.16 | 324.14 |
| Shellcode | 7.46 | 9.92 | 9.08 | 280.76 |
| Worms | 7.17 | 11.28 | 13.18 | 256.61 |

通过对比发现，本文方法消耗的时间比 SMOTE 方法及 Borderline-SMOTE 方法消耗的时间长，这是由于本文算法需要耗费大量时间来进行机器学习训练，同时在 3 种方法的数据传递中涉及较多的文件读取操作，因此本文算法相比传统方法耗时较长。

**结束语** 为了提升针对不平衡样本的异常流量检测效果，本文使用近年来在图像生成领域广泛使用的变分自编码器模型，将该模型应用于异常流量样本生成领域，提出了一种基于变分自编码器的少量样本生成方法。在使用变分自编码器进行少量样本生成前，首先，基于 KNN 算法设计了 KNN 少量样本选择算法对原始少量样本进行初步筛选，筛选出部分需要扩充的少量样本集；接着，基于 DBSCAN 算法设计了 DBSCAN 类内聚类算法，用于对 KNN 算法筛选后的部分少量样本进行聚类处理，分别保存聚类后不同簇类的少量样本；最后，利用变分自编码器对不同簇类的少量样本分别进行训练，以建立概率密度分布模型，并按照概率密度分布模型分别生成样本，将生成的样本与原始训练样本合并，送入决策树分类器进行训练，测试。本文采用 UNSW-NB15 数据集来评估本文方法的有效性。通过对比原始样本与 SMOTE 方法生成的样本，本文方法生成的样本有效提升了决策树分类器的分类效果，多数情况下本文算法的召回率、$F1$ 分数指标均高于原始样本及传统方法。

本文还存在以下几方面的不足：首先，在少数类样本数量很少的情况下，本文经过 KNN，DBSCAN 算法筛选后的少量样本数量会继续减少，无法保证样本数量，在未来研究中可以考虑不同上采样方法的融合，在执行本文的数据筛选算法前，通过其他类型的上采样算法对数据进行扩充；其次，当少数类异常样本的数量较少时，如果这类异常样本不具备普遍性，则经过包括本文方法在内的上采样方法生成的新样本仍不具备普遍性，最终训练的预测模型虽然效果有所提升但仍无法达到较高的准确度；接着，目前本文设计实现的模型均为单线程处理，相比传统方法时间消耗增长很大，后续的改进可以引入多线程方式，同时通过优化减少文件操作步骤来减少时间消耗；最后，本文采用深度学习生成模型中的变分自编码器模型，并未尝试其他生成模型及如何改进变分自编码器模型以提升效果。在未来的研究中，将考虑采用如生成对抗网络(GAN)等其他深度生成模型，同时合理改进变分自编码器模型以进一步提高样本的生成质量，提升机器学习分类器的检测效果。

## 参 考 文 献


[1] China Internet Network Information Center. The 44th statistical report on the development of Internet in China[J]. Internet World, 2019(10): 74-91.

[2] ZHANG Y Q, ZHOU W, PENG A N. Overview of Internet of things security[J]. Computer Research and Development, 2017, 54(10): 2130-2143.

[3] GUI C N. Global Internet of things attacks increased by 280% in the first half of 2017[J]. China Information Security, 2017(9): 10.

[4] ZHAO X. Design and implementation of network traffic detection system[D]. Northeast Normal University, 2011.

[5] ZHANG Y Q, DONG Y, LIU C Y, et al. Current situation, trend and Prospect of deep learning application in Cyberspace Security[J]. Computer Research and Development, 2018, 55 (6): 1117-1142.

[6] KANG S L, FAN X P, LIU L, et al. Research on P2P Botnets Detection Based on the ENN-ADASYN-SVM Classification Algorithm[J]. Journal of Chinese Computer Systems, 2016, 37(2): 216-220.

[7] MO Z, GAI Y R, FAN G L. Credit card fraud classification based on GAN-AdaBoost-DT imbalanced classification algorithm[J]. Journal of Computer Applications, 2019, 39(2): 618-622.

[8] KIM J H. Time Frequency Image and Artificial Neural Network Based Classification of Impact Noise for Machine Fault Diagnosis[J]. International Journal of Precision Engineering and Manu-

facturing,2018,19(6):821-827.
[9] PUN J,LAWRYSHYN Y. Improving Credit Card Fraud Detection using a Meta-Classification Strategy[J]. International Journal of Computer Applications,2012,56(10):41-46.
[10] CHAWLA N V,BOWYER K W,HALL L O,et al. SMOTE:Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research,2002,16(1):321-357.
[11] HE H,BAI Y,GARCIA E A,et al. ADASYN:Adaptive Synthetic Sampling Approach for Imbalanced Learning[C]// IEEE International Joint Conference on Neural Networks (IJCNN 2008). IEEE,2008.
[12] HAN H,WANG W Y,MAO B H. Borderline-SMOTE:A New Over-Sampling Method in Imbalanced Data Sets Learning[C]// International Conference on Intelligent Computing. Berlin,Heidelberg:Springer,2005:878-887.
[13] ZHU T,LIN Y,LIU Y. Synthetic minority oversampling technique for multiclass imbalance problems[J]. Pattern Recognition,2017,72:327-340.
[14] DOUZAS G,BACAO F. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE[J]. Information Sciences,2019,501:118-135.
[15] CASTRO C L,BRAGA A P. Novel Cost-Sensitive Approach to Improve the Multilayer Perceptron Performance on Imbalanced Data[J]. IEEE Transactions on Neural Networks and Learning Systems,2013,24(6):888-899.
[16] LI Y,LIU Z D,ZHANG H J. Overview of integrated classification algorithm for unbalanced data[J]. Computer Application Research,2014,31(5):1287-1291.
[17] GALAR M,FERNANDEZ A,BARRENECHEA E,et al. A Review on Ensembles for the Class Imbalance Problem:Bagging-,Boosting-,and Hybrid-Based Approaches[J]. IEEE Transactions on Systems,Man and Cybernetics,Part C (Applications and Reviews),2012,42(4):463-484.
[18] SHI J R,MA Y Y. Research progress and development of deep learning[J]. Computer Engineering and Application,2018,905(10):6-15.
[19] KINGMA D P,WELLING M. Auto-Encoding Variational Bayes[J]. arXiv:1312.6114.2013.
[20] LIU F. Research on the theory and application of deep self encoder[D]. Wuxi:Jiangnan University,2018.
[21] MA H Q,MA S P,XU Y L,et al. Image denoising[J]. Computer Engineering and Application,2018,54(4):199-204,236.
[22] YIN B C,WANG W T,WANG L C. A review of deep learning research[J]. Journal of Beijing University of Technology,2015(1):48-59.
[23] ZENG X Y,YANG Y,WANG S Y,et al. A hybrid recommendation algorithm based on deep learning[J]. Computer Science,2019,46(1):126-130.
[24] LIU S,HUANG Y,HU J,et al. Learning local responses of facial landmarks with conditional variational auto-encoder for face alignment[C]// 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE,2017:947-952.
[25] OSADA G,OMOTE K,NISHIDE T. Network intrusion detection based on semi-supervised variational auto-encoder[C]// European Symposium on Research in Computer Security. Cham:Springer,2017:344-361.
[26] ZHAI Z L,LIANG Z M,ZHOU W,et al. Review of variational self encoder models[J]. Computer Engineering and Application,2019,55(3):1-9.
[27] MOUSTAFA N,SLAY J. UNSW-NB15:a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)[C]// 2015 Military Communications and Information Systems Conference (MilCIS). IEEE,2015.

**ZHANG Ren-jie**, born in 1995, M. S. candidate, is a student member of China Computer Federation. His main research interests include network security, machine learning.

**CHEN Wei**, born in 1979, Ph.D, professor, is a member of China Computer Federation. His main research interests include wireless network security, mobile Internet security.