融合集成学习技术和PSO-GA算法的特征提取技术的入侵检测方法

王 军，司昌馥，王凯鹏，付 强

（沈阳化工大学 计算机科学与技术学院，沈阳 110142）

摘 要：针对工业网络的安全问题，提出了一种新的入侵检测方法，方法的具体创新之处分为两点。首先，在处理数据特征过程中，针对原始数据维度较高的问题，提出一种参数动态调整的粒子群优化-遗传混合算法，用于特征提取，成功筛选出了对模型训练有意义的特征子集，加快了模型训练速度。其次，在构建机器学习模型时，使用了堆叠集成学习框架对多个模型的输出结果进行泛化，以获得整体预测精度的提升。共在两个数据集上验证了本文方法的检测性能，试验结果表明：在公开的入侵检测数据集CICDS-2017上的检测精确度达到了95%，在由美国密西西比州立大学的Lan Turnipseed从天然气管道控制系统收集到的真实工业数据集上也达到了93%的精确度。

关键词：计算机应用；工业控制系统；入侵检测；集成学习；特征提取

中图分类号：TP399 **文献标志码**：A **文章编号**：1671-5497(2025)04-1396-10

DOI：10.13229/j.cnki.jdxbgxb.20230751

Intrusion detection method based on ensemble learning and feature selection by PSO-GA

WANG Jun，SI Chang-fu，WANG Kai-peng，FU Qiang

（*College of Computer Science and Technology*，*Shenyang University of Chemical Technology*，*Shenyang* 110142，*China*）

Abstract：In response to the security issues in industrial networks，a new intrusion detection method is proposed. The specific innovations of the method are divided into two aspects. First，in the process of processing，in order to solve the problem of high dimensionality of the original data，a particle swarm optimize genetic algorithm（PSO-GA）hybrid algorithm with dynamically adjusted parameters was proposed for feature extraction. It successfully screened out a subset of features that are meaningful to model training and accelerated training speed. Secondly，when building a machine learning model，theStacking integrated learning framework is used to generalize the output results of multiple models to improve the overall prediction accuracy. The experimental results on both two datasets show that the detection precision on the publicly available intrusion detection dataset CICDS-2017 has reached 95%，and

收稿日期：2023-07-17.

基金项目：辽宁省自然科学基金项目(2022-MS-291)；国家外国专家项目(G2022006008L)；辽宁省教育厅高校基本科研项目(LJKMZ20220781，LJKMZ20220783，LJKQZ20222457).

作者简介：王军(1978-)，男，教授，博士.研究方向：工业网络安全.E-mail：wj_software@hotmail.com

通信作者：付强(1990-)，男，讲师，博士.研究方向：工业网络安全.E-mail：qiang.fu@outlook.com

it also has a 93% precision on a real industrial dataset developed by Lan Turnipseed from the gas pipeline control system.

Key words: computer application; industrial control system; intrusion detection; ensemble learning; feature selection

0　引　言

传统的工业控制系统(ICS)是基于物理隔离的，因此，人们更加关注系统的功能安全问题而忽视了系统的信息安全问题。然而今天的ICS正在向信息化方向发展，随着工业互联网的发展，ICS逐渐成为一个开放的互联网系统，针对其的网络攻击也层出不穷。工业控制系统的网络安全问题日益严峻，工业网络安全问题受到了比传统功能安全更多的关注和研究[1-3]。传统针对工业网络的安全方法主要依赖部署在边缘的安全设备，例如可以根据自定义的规则过滤网络流量的防火墙。然而，这种被动的防御策略只能拦截已知的网络攻击，在面对新的网络威胁时无法做到有效识别和拦截，因此，需要更加主动、智能的入侵检测方法。

机器学习方法作为一种快速发展的人工智能方法，可以从大量已知数据中挖掘出隐藏的规律，并利用学到的规律对新数据进行预测和分类。随着人工智能的发展和普及，机器学习在各种学科和领域都扮演着重要的角色，在入侵检测领域也不例外。结合机器学习方法的入侵检测系统可以通过训练学习已知的网络攻击，进而识别出以前没有出现过的新的异常网络流量，大大提高了检测精度。本文提出的方法是结合集成学习技术和特征选择技术来提高对网络攻击的检测精度并且在公共CICIDS-2017数据集和真实天然气行业数据集上进行测试，结果证明了本文此方法的高效率和高精确性。

1　相关工作

1.1　集成学习方法

机器学习在构建入侵检测模型中发挥着重要作用。随着大数据时代的到来，数据的规模不断扩大且维度不断增长，单个分类器对样本的预测能力变得有限，研究适当的集成方法对构建有效的机器学习入侵检测模型至关重要。2015年，Gaikwad等[1]提出了一种基于机器学习集成方法的入侵检测技术。采用以REPTree为基类的集成Bagging方法来实现入侵检测系统(Intrusion detection system，IDS)。选择NSL_KDD数据集进行训练和测试。试验结果表明：具有REPTree基类的Bagging集成表现出最高的分类精度且建模所需的时间更少。2018年，Shen等[2]提出了一种使用随机子空间的集成方法，其中选择极限学习机(Extreme learning machine，ELM)作为基分类器。为了优化集成模型，提出了一种基于蝙蝠算法(BA)的集成剪枝方法。同时，在BA中定义了适应度函数，以获得改进的分类器子集。实证结果表明：基于随机子空间的集成方法在准确性和鲁棒性方面优于单独使用的ELM，同时节省大量计算资源。2021年，Bhati等[3]提出了一种使用XGBoost的集成IDS，表明使用XGBoost的集成IDS可以获得很好的检测结果，试验在KDD-Cup99数据集上进行，准确率高达99.95%。

1.2　基于改进粒子群优化算法的特征提取技术

标准粒子群优化(PSO)算法是常用的参数优化方法，速度快，但也容易陷入局部最优，造成停滞问题[4]。2015年，Ahmad等[5]分析了主成分分析(PCA)、遗传算法(GA)等进化优化方法用于特征提取的效果和优缺点。PCA把特征首先投影到主空间中，然后根据特征值选择特征，但具有最高特征值的特征可能无法保证为分类器提供最佳灵敏度。遗传算法(GA)等优化方法也已被用来搜索变换特征中最具辨别力的子集。Ahmad等[5]提出了一种基于PSO的特征子集选择，与GA相比，有更好的性能。Dickson等[6]在2020年提出了一种多目标粒子群优化方法，使用NSL-KDD数据集和5种不同的机器学习分类器，分析了粒子群算法对模型的优化作用。Aziz等[7]在2022年提出了一种新的包装器特征选择模型，称为恢复粒子群优化(RPSO)，以选择高度相关的特征数据，同时考虑到最优特征迭代过程中出现的停滞问题，利用随机值来克服停滞问题，减少数据量并缩短处理时间。用随机森林算法对选择出的特征进行分类。使用NSL-KDD基准数据集来

评估RPSO方法的效果。试验表明:与标准PSO算法相比,准确率由83%提高到了85%。Wei等[8]提出了一种基于深度信念网络(DBN-IDS)的入侵检测模型。利用GA优化来降低数据集的维度。基于学习因子和自适应权重,利用PSO算法计算GA算法的初始值,从而找到初始解,文献[8]中的模型在完整版和缩小版的NSL-KDD数据集上的准确率分别达到了82.36%和66.25%。

综上所述,现有的一些利用集成学习方法和改进PSO算法参数优化的入侵检测方法取得了较高检测精度,在对PSO算法进行改进时可以从两方面入手:一方面对基础PSO算法的参数进行改进和优化;另一方面可以引入其他算法,比如遗传算法(GA)作为PSO算法的算子,提高PSO算法的搜索能力。目前,大多数性能测试仍然是在非工业入侵检测数据集上进行的,依然无法证明上述方法可以在真实的工业控制系统上使用。针对这个问题,本文将在两个数据集上进行性能测试,以确保本文方法在现实工业环境中仍然是有用且高效的。

2 数据集

2.1 CICIDS-2017数据集

自CICIDS-2017数据集被开发出来以后,其吸引了大量研究人员和学生的关注,基于对该数据集的分析和研究,许多新的入侵检测模型被提出。该数据集来自加拿大网络安全研究所,由8个不同的文件组成,其中包含了5 d的正常和攻击流量数据[8]。

这个数据集很大,同时数据集中存在严重的类别不平衡问题,这意味着大部分训练时间都用于学习如何检测数量较多的类别,对数量较少的类别则无法有效学习和识别。这会导致模型的检测结果有较低的精确度以及较高的误报率[9]。

为了解决这些问题,本文参考了Goryunov[10]、Stiawane[11]、Salo[12]等提出的方法,通过合并少数类别以形成新的攻击类别并删除多数类别中的部分数据,从而有效地缓解了类别不平衡。预处理后的数据集特征可称为CICIDS2017_sample,新数据集的描述如表1所示。

2.2 天然气管道数据集

除了使用CICIDS-2017数据集外,为了确保本文方法在真实工业互联网环境中仍然能够发挥作用,本文还使用了真实工业环境中收集的数据集。该数据集是美国密西西比州立大学研究团队从天然气管道获得的网络流量数据日志[13]。天然气管道数据日志是在试验室捕获的,包括正常数据和各种网络攻击数据。具体数据分布如表2所示。

表1 简化后的数据集

Table 1 CICIDS2017_sample dataset

编号	标签	样本数量/个
1	BENIGN	22 767
2	DoS	19 035
3	PortScan	7 946
4	BruteForce	2 767
5	WebAttack	2 180
6	Bot	1 966

表2 天然气管道数据集的描述

Table 2 Description of gas pipeline dataset

标签	缩写(编号)
Normal	Normal(0)
Naïve Malicious Response Injection	NMRI(1)
Complex Malicious Response Injection	CMRI(2)
Malicious State Command Injection	MSCI(3)
Malicious Parameter Command Injection	MPCI(4)
Malicious Function Code Injection	MFCI(5)
Denial of Service	DOS(6)
Reconnaissance	Recon(7)

3 入侵检测网络模型

3.1 数据预处理

3.1.1 特征归一化

基于参数的模型和基于距离的模型都需要特征归一化。对连续特征进行归一化常用的方法有两种:线性归一化和标准差归一化。本文采用线性归一化方法,将数据在0和1之间映射。其公式如(1)所示:

$$x_{\_scale}=\frac{x-x_{\min}}{x-x_{\max}} \tag{1}$$

式中:$x_{_scale}$为归一化处理之后得到的值;x为数据集中的初始值;$x_{\max}$和$x_{\min}$为要归一化的值的最大边界和最小边界。

3.1.2 连续型特征归一化

如图1所示,CICIDS-2017_sample数据集中数据不平衡的问题虽然有所缓解但是仍然存在,如果直接使用原始数据进行训练,会导致检测模型倾向于多数类。因此,检测性能会很差,误报率

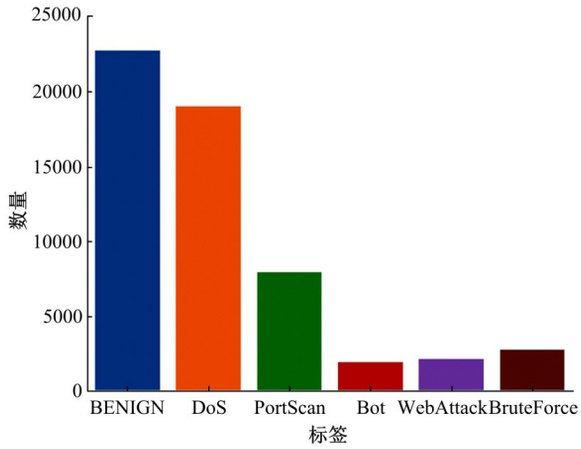

图 1　CICIDS-2017_sample 标签分布

Fig.1　Labels for CICIDS-2017_sample

升高，准确率下降。针对这个问题，本文采用了合成少数类过采样（SMOTE）算法来进一步解决数据不平衡问题。

SMOTE算法是常用的采样方法之一。简而言之，该算法的思想就是随机创建一些新的少数类样本，使这些少数类的数量不断增加。具体来说，对每个少数类的样本数据，找到与其最近邻的同类数据，然后在它们之间进行插值以形成新的数据。这样，创造足够多的新数据，少数类样本的数据量增加从而让不同类别的数据量趋于平衡[14]。

3.2　模型选择

入侵检测是一个二分类或多分类问题，而机器学习方法在此类问题上一直有出色的性能。其中一种机器学习方法是基于树结构的机器学习。例如决策树、随机森林、XGBoost等。本文在使用上述机器学习方法的基础上，使用集成学习技术把上述树形的分类器作为基础分类器集成到一起形成新的集成学习模型。不仅提高了检测精度，还加快了检测速度。决策树（DT）是一种树形结构，自上而下地对样本的属性进行判断，直到到达叶子节点并输出最终结果[15]。在实际应用中，由于其结构简单，DT常常被用于其他算法，例如随机森林（RF）的基础模型。RF是基于投票原理的分类器，通过多个决策树的预测结果选择投票率最高的最终结果作为分类输出[16]。XGBoost也是另一种结合了许多决策树的机器学习算法，为了提高其速度和性能，XGBoost中还使用了梯度下降算法[17]。

还有其他的机器学习方法可以应用于多分类问题，比如K-NN、SVM等。而且它们在时间复杂度方面的表现并不逊色于本文使用的方法。比如K-近邻算法的复杂度为$O(N\times P)$，而决策树的算法复杂度为$O(N\times N\times P)$。不过本文使用的方法都支持多线程，可以成倍数地减少训练时间，提高训练效率。

选定了基础模型后，为了获得更好的性能需要对模型的超参数进行调整。对于DT算法，使用基尼指数，即CART算法，然后建立模型。比使用信息增益理论构建ID3树的开销更少，性能更好。S为所有子树的集合，CART在S中选择使式（2）最小化的树：

$$C(S)=\hat{L}n(S)+\alpha|S| \tag{2}$$

式中：$|S|$为树的基数；α为常数；$\hat{L}n(S)$为使用树S的经验风险。

由于更深的树有更多的子树，因此，树的深度D是CART算法的一个重要参数。而对RF和XGboost而言，由于它们的结果是基于许多决策树的多数表决结果，因此，决策树的数量T是影响性能的一个重要参数。

为了选出模型的最佳参数，本文采用了网格搜索法[18]。此算法的执行过程描述如下：首先把树的数量（T）及深度（D）这两个亟须确认的参数设为1，以此为起点开始进行尝试。慢慢增加这两个参数的数值，直到训练的精度不再呈现上升趋势。最后，将D和T设为适当值。从图2可以看到树深度对准确性的影响。树的深度在增加到5层后继续增加则对预测精度的提升极其有限。考虑到越来越多的层数还会明显延长模型运行时间，降低模型的检测速度，由此可以得出树的最佳

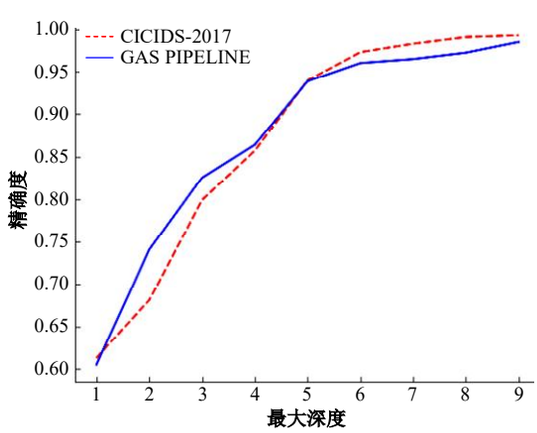

图 2　树的深度对精确度的影响

Fig. 2　Impact of tree depth on precision

深度为5。同理,其他参数也可以使用该方法进行测试和调整,最终确定最合适的参数值。

3.3 PSO-GA算法优化的特征提取方法

基础的PSO算法利用一组粒子在搜索区域中的移动以寻找最佳参数在搜索空间中的位置。该算法首先随机创建粒子群,并在空间中逐步更新每个粒子的位置,每次迭代过程中单个粒子与它本身的经验以及其他粒子的经验进行比较来更新其下一步的运动方向和速度。所谓的经验就是用适应度函数来评估每一个粒子的参数是否更优。第一次比较是将粒子当前的位置与其自身经历过的最佳位置进行比较,第二次比较是将粒子与群体中所有粒子的最优粒子位置进行比较。这两个比较的对象可以分别称为个人最佳(p-best)和全局最佳(g-best)[19]。

在每次迭代中,粒子运动的方向由它先前的最佳粒子位置和全局最佳位置共同决定,速度的计算也与方向同理。具体来说,就是综合考虑粒子的当前速度、粒子与最佳位置之间的距离、处于全局最佳位置粒子的速度这3个因素,从而得出下一步迭代的合理速度。获得新的速度和方向后,就可以确定新的粒子位置。该过程将连续执行,直到找到最优的参数或达到最大的迭代次数。

式(3)用于更新PSO算法中的粒子速度:

$$V_{ij}(t+1)=V_{ij}(t)+C_1R_1[V_{ij}(t)-X_{ij}(t)]+C_2R_2[P_{gj}(t)-X_{ij}(t)] \tag{3}$$

在[0,1]范围内随机选取两个随机变量,称为R_1和R_2;C_1和C_2是取正值的两个常数参数;$V_{ij}(t)$和$X_{ij}(t)$表示当处在迭代t时的粒子速度和粒子位置。

式(4)用于更新PSO算法中的粒子位置:

$$X_{ij}(t+1)=X_{ij}(t)+V_{ij}(t+1) \tag{4}$$

作为一种随机搜索算法,PSO算法具有很大的局限性——粒子会过早陷入局部最优解。因此,虽然PSO算法优化参数的速度比其他算法更快,但优化参数的能力或者说全局寻优能力不强。为了尽可能避免种群陷入局部最优解的陷阱,需要对基础的PSO算法进行改进,增加执行过程中粒子的随机性。

首先,可以在式(3)和(4)中引入惯性权重w,从而得到式(5)和(6):

$$V_{ij}(t+1)=wV_{ij}(t)+C_1R_1[V_{ij}(t)-X_{ij}(t)]+C_2R_2[P_{gj}(t)-X_{ij}(t)] \tag{5}$$

$$X_{ij}(t+1)=X_{ij}(t)+V_{ij}(t+1) \tag{6}$$

式中:w决定了当前速度对粒子速度变化的影响程度,w值越大,粒子的速度就越大,粒子的全局搜索能力越强;较小的w值则会让速度变慢,有利于粒子在局部位置搜索和优化。

另外,在搜索过程中,可以对w的值进行动态调整,即在迭代过程中线性地减小w的值,具体的参数调整方法可表达为式(7):

$$w=w_s-(w_s-w_e)\frac{t}{T_{\max}} \tag{7}$$

式中:w_s为初始惯性权重;w_e为可取的最小权重;$T_{\max}$表示最多允许进行多少次迭代;T表示当前已经执行了多少次迭代。

基于同样的思路,还可以对式(2)中的常数参数C_1、C_2进行相似操作。在算法执行的早期设置较大的C_1值和较小的C_2值,这样的好处是在搜索初期个体粒子的搜索经验占主导地位,更有可能遍历整个搜索空间,避免了快速陷入某个局部最优解。具体实现方法为对参数C_1、C_2进行线性的动态调整,其公式为:

$$C_1=R_3+\frac{R_4}{T_{\max}}t \tag{8}$$

$$C_2=R_5-\frac{R_6}{T_{\max}}t \tag{9}$$

式中:R_3、R_4、R_5和R_6为初始值。通过模拟测试函数,$0<R_3+R_4\leqslant 2$,$R_5-R_6\geqslant 1$,该算法尽可能避免了局部搜索的缺点,加快了收敛速度。

改变参数的计算方式只是多种改进粒子群优化算法的思路之一,为了进一步对PSO算法进行改进,还可以引入新的算法实现进一步的改进。遗传算法(GA)采用了选择、交叉等参数优化方法,增强了全局寻优能力,提高了参数的收敛精度。但这种机制也导致了算法收敛时间增加,可见遗传算法并不是一种好的方法。然而,如果将遗传算法引入PSO算法中,形成PSO-GA混合算法[20],遗传算法充当粒子群算法的算子,当种群陷入局部最优解时通过遗传算法使粒子发生变异,从而跳出局部最优解,增强粒子群的搜索能力。这种方式可以发挥两种算法的优点。算法过程简述如下:

(1)设置最大迭代次数,初始化种群中的每个随机粒子,为该群体中每个粒子设置初始位置和速度。

(2)获取每个粒子的适应度值并初始化$p_{_best}$和$g_{_best}$。

(3)根据$p_{_best}$和$g_{_best}$更新粒子的速度和位置。

(4)根据新的粒子适应度更新$p_{_best}$和$g_{_best}$。

(5)判断粒子是否陷入局部最优。如果是，则执行遗传算法对种群进行交叉和变异操作。

(6)验证是否达到结束条件。如果当前迭代达到最大次数或者找到最优参数，则停止迭代并输出最优解。

图3为粒子的进化过程，从折线图的走势可看出，在迭代过程中，粒子的全局搜索能力较强，没有出现短时间就停止搜索，陷入局部最优解的糟糕情况。迭代80次左右粒子接近全局最优解，考虑到PSO算法迭代速度快但容易停滞不前的特点，可以说，本文算法通过对基础PSO算法进行动态参数调整，并引入遗传算法，增加全局搜索能力的策略是非常有针对性的，成功改善了基础PSO算法的问题，使寻找参数的效率大大提高。

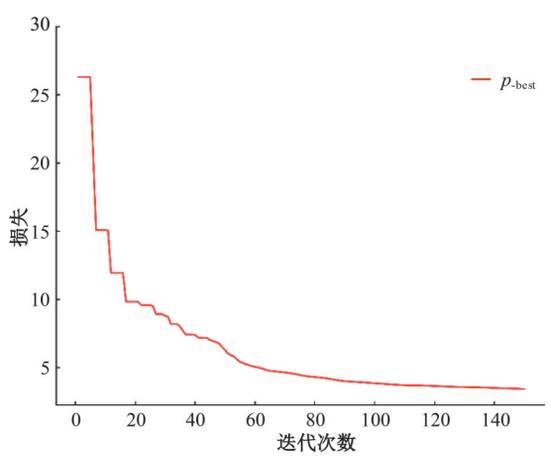

图3　损失随着迭代次数下降

Fig. 3　Cost decreases with number of iterations

需要处理高维度的数据集时，由于原始数据集中包含大量对模型训练无意义的冗余特征，所以提取出高相关性特征是所有特征提取算法的共同目标。为了达到这一目标，主要有两种选取特征的思路：过滤法(Filter method)和包装法(Wrapper method)。过滤器方法注重分析特征向量的数学规律，通过统计学的指标对每个特征进行单独的打分并排序，但是这种方法不会考虑机器学习模型的工作过程，且主要适用于互相独立的特征之间。包装法选择特征时则会将预测模型纳入考虑的范围。将不同的特征子集用来训练不同的模型，训练得到的模型的泛化性能可以为该子集评分。

图4显示了在CICIDS-2017数据集上不同特征子集经过训练后得到的准确率，从图中可以发现包装方法是计算密集型的，耗时更长，但可以针对特定的模型提供相应表现最佳的特征集。由于PSO算法拥有迭代速度快的优势，因此，格外适用于包装法。本文提出的参数动态特征调整的PSO-GA算法，不仅保留了PSO算法的速度优势，还增强了算法的搜索能力。图4展示了不同特征子集在模型上的训练准确率，原始的CICIDS-2017数据集包含79个特征，从图4可以看到，在中间最上方的红色椭圆处，标记出了所有特征子集中检测效果最佳的一个子集，因此，只要在79个原始数据特征中保留39个重要的特征即可达到95%以上的准确率。

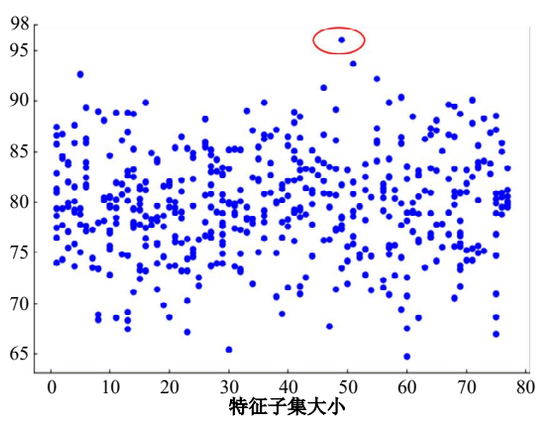

图4　不同特征子集对应的准确率

Fig. 4　Accuracy corresponding to different feature subsets

3.4　集成学习方法

集成学习是通过集成基础模型来解决分类问题的方法。每个基本模型都会做出自己的标签预测，然后集成学习模型最终的类别标签预测由投票产生[21]。因此，集成学习比单一的基础分类器具有更高的精度，并且通常来说集成学习模型的泛化能力也比基本模型强得多。多种集成学习方法已经被提出和使用，例如简单平均、加权平均、多数投票、加权投票、集成堆叠等[20]。其中一些方法仅由一种基础模型组成，而另一些方法则是基于不同类型分类器的组合。本文使用堆叠集成方法来集成3.2节中提到的基于树的机器学习方法，以提高检测性能。

从图5可以看出，堆叠集成模型[22]由两层组

成。一系列不同的基本模型在第一层工作,同时在下一层有一个元分类器,它从第一层获取预测作为输入,然后生成最终结果。

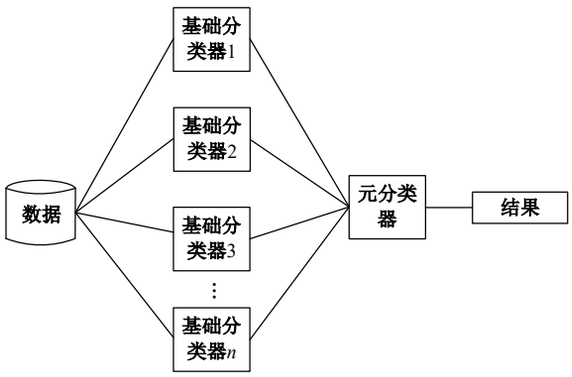

图5 堆叠集成

Fig. 5 Stacking Ensemble

3.5 模型框架图

图6为本文模型的执行流程。首先,收集数据,包括正常数据和多种类型的攻击;其次,对数据进行预处理归一化,如果数据集存在类别不平衡的问题,则执行SMOTE算法生成新的少数类数据;下一阶段,基于PSO-GA参数优化算法,进行特征选择,降低模型检测的计算成本,提高检测效率;然后,使用基本模型对数据集分别进行预测;最后一步,所有基本模型的预测将作为堆叠集成模型的输入,由集成模型输出最终的分类结果。

4 性能分析

本试验使用两个数据集来测试所提出的集成模型检测方法的性能,即CICIDS-2017数据集和天然气管道数据集。首先,在CICIDS-2017数据集上,比较了单一树形模型、未提取特征的集成学习模型和特征提取后的集成模型三者的检测性能,以证明特征提取方法和集成学习模型对检测性能的改进效果。然后将本文方法与其他文献中提出的检测方法进行比较。最后,将本文方法应用于真实的天然气管道数据集,证明此方法能够保证对工业系统中真实数据的良好检测性能。

4.1 评价指标

该试验使用了多分类混淆矩阵(见表3)来展示模型的实际检测效果[23]。混淆矩阵可以显示出实际值和预测值之间的差异,在矩阵中可以清楚看到对多少数据的判别是正确的,又有多少数据没有被正确地分类。对多分类任务,它的混淆矩阵是一个$N\times N$的矩阵,其中N代表的是数据

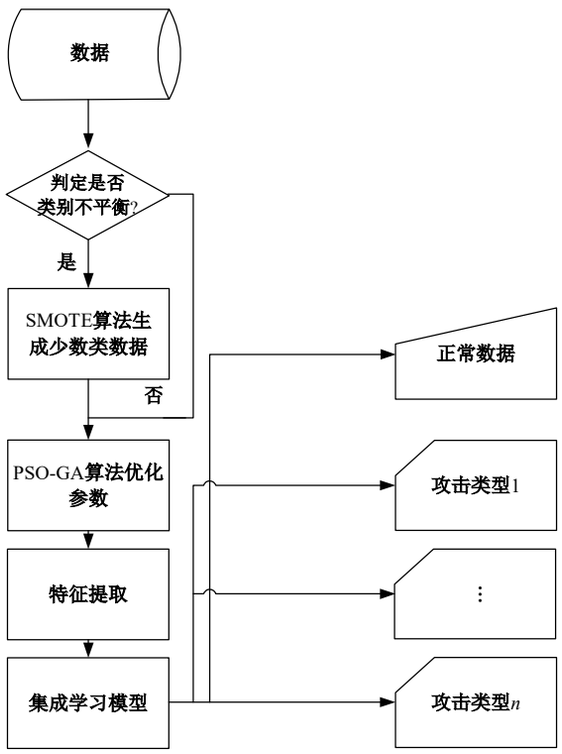

图6 本文模型流程图

Fig. 6 Flow chart of proposed model

类别的数量。

表3 多分类任务混淆矩阵

Table 3 Confusion matrix for multi classification tasks

	类别1	类别2	…	类别n
类别1	A_{11}	A_{12}	…	A_{1n}
类别2	A_{21}	A_{22}	…	A_{2n}
⋮	⋮	⋮		⋮…
类别n	A_{n1}	A_{n2}	…	A_{nn}

根据混淆矩阵,可以计算出以下3个评价模型的参数指标:

Precision(精确度):它表示所有模型认为类别A_1的样本中确实属于类别A_1的样本所占的比例。

Recall(召回率:)它表示有多少实际上是类别A_1的样本被预测为类别A_1的。

F_1分数:它是上面两个参数的调和平均值。

$$\text{Percision}(A_1)=\frac{A_{11}}{\sum_{i=1}^{n}A_{i1}} \tag{10}$$

$$\text{Recall}(A_1)=\frac{A_{11}}{\sum_{i=1}^{n}A_{1i}} \tag{11}$$

$$F_1\text{分数}=\frac{2\times\text{Percision}\times\text{Recall}}{\text{Percision}+\text{Recall}} \tag{12}$$

4.2　在CICIDS-2017数据集上的性能分析

4.2.1　单一模型与集成模型的对比

从表4可以看出，集成学习方法与单一的基础模型相比具有明显的优势。集成学习方法在用于评估模型性能的三指标数据上都有一定的提升。只有决策树模型在检测时间上对集成模型具有优势。但考虑到单一决策树模型在精确率、召回率和F_1分数上均逊于集成方法，因此，可以得出结论：集成方法在综合性能上优于所有单一模型方法；在同样使用集成学习模型进行检测的情况下，如果在检测前对特征集合进行特征提取的处理，则模型在精确度、召回率和F_1分数上也会有小幅提升。同时，特征提取后可以大大加快检测速度，与不进行特征提取的情况相比，可以看出，经过特征提取处理后，同一个模型达到类似甚至稍微更好的检测结果只需一半的时间。

表4　在集成和特征选择处理前后的比较

Table 4　Comparison before and after ensemble and feature selection

方　法	精确度	召回率	F_1分数	CPU时间/s
决策树	0.925	0.964	0.951	22.8
随机森林	0.862	0.996	0.906	3 min 5
极端梯度提升	0.903	0.997	0.935	4 min 14
堆叠集成	0.936	0.975	0.912	3 min 25
堆叠集成加特征提取	0.951	0.987	0.953	1 min 30

4.2.2　集成模型与其他机器学习方法的对比

表5展示了本文方法与其他文献提出的预测方法之间的检测效果的比较。这些用于对比的文献使用了多层感知机（MLP）、长短期记忆网络（LSTM）、基于图的深度特征学习（DeepGFL）等机器学习方法。与文献[24]和[25]相比，本文方法在精确度、召回率、和F_1分数上均有领先。另一篇参考文献[26]中提出的方法得出的评价指标比本文方法更好，但文献[26]中的模型适用于二分类问题，只能判断出数据为正常数据或攻击数据，无法针对更多种不同的攻击类型进行多分类处理，而本文方法可以对数据集中的多个不同类别数据进行多分类。

表5 与其他方法的比较

Table 5　Comparison with other method

来 源	方 法	精确度	召回率	F_1分数	类别数
本文	特征提取加堆叠集成	0.951	0.987	0.953	6
文献[24]	MLP	0.871	0.995	0.873	6
文献[25]	DeepGFL	0.948	448	0.531	12
文献[26]	MLP	0.884	0.862	0.872	2
文献[26]	LSTM	0.984	0.898	0.895	2

4.3　在天然气管道数据集上的性能分析

上一节的试验数据主要是在CICIDS-2017数据集上得出的，初步证明了该方法的检测性能，为了进一步证明该方法在真实工业互联网入侵检测任务中仍然有效，本节将相同检测模型应用到真实工业系统中采集的天然气管道数据集上。图7、表6和表7展示了模型的检测结果。从图7、表6、7中可以看出，虽然本文方法在个别攻击类别上不能保证高精确度，但是在完整数据集上的平均表现良好。

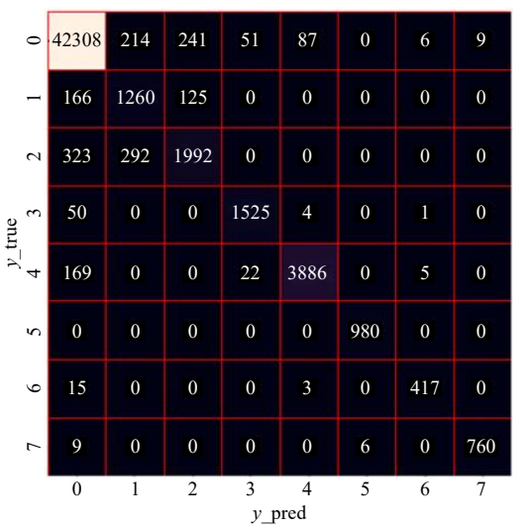

图7　分类的混淆矩阵

Fig. 7　Confusion matrix of classification

表6　每一类数据的精确、召回率、F_1分数

Table 6　Precision, recall, F_1-score of every kind of data

类别	精确度	召回率	F_1分数
0	0.98	0.99	0.98
1	0.72	0.83	0.77
2	0.85	0.77	0.81
3	0.96	0.96	0.96
4	0.97	0.94	0.96
5	0.99	1.00	1.00
6	0.96	0.95	0.95
7	0.99	0.98	0.98

表7　整个数据集上的平均结果

Table 7　Average results of the whole dataset

	精确度	召回率	F_1分数
宏平均	0.93	0.93	0.93
加权平均	0.97	0.97	0.97

表6显示了模型对天然气管道数据集中的8类数据——其中包括1种正常数据以及7种攻击数据的检测结果。从表5中可以看到,经过特征提取过程的集成学习模型的表现在大多数类别的数据上仍然优秀。除NMRI和CMRI这两类的数据外,对其他各类数据的分类精确度均达到96%以上。同时,如果直接考察本模型在完整数据集上的总体表现则如表7所示,所有8种数据的预测精确度的算数平均值可以达到93%,进一步考虑每种数据的权重,加权后的平均精确度将高达97%。

图7为对天然气管道数据集进行分类后得到的多分类混淆矩阵,总结了具体的检测结果。从混淆矩阵上可以更直观、更清楚地看到模型对各类数据的预测和分类情况。由图7可知,尽管大约有2000个样本没有被分到正确的类别中,但显然大部分数据都是正确分类的。因此,可以得出这样的结论:本文方法在实际的工业控制系统中也是有效的。

5 结 论

(1)使用集成学习技术构建机器学习模型,可使模型的学习能力更强。

(2)分析论证了在集成学习模型中使用树形基本分类器的必要性,并通过网格搜索法确定了模型的各个参数。

(3)使用参数动态调整的粒子群优化算法进行特征提取,提高模型学习效率。对粒子群优化算法的参数进行动态调整,提高算法效率。

(4)对PSO算法本身进行改进后,又在参数寻优过程中引入GA算法,使用PSO-GA混合算法进一步避免了PSO算法陷入局部最优解的情况,提高了PSO算法的参数搜索能力。

(5)在试验中使用了真实的工业数据,证明了此方法在真实工业互联网中可以发挥入侵检测的作用。

参考文献:

[1] Gaikwad D P, Thool R C. Intrusion detection system using bagging ensemble method of machine learning [C]//International Conference on Computing Communication Control and Automation, Pune, India, 2015: 291-295.

[2] Shen Y, Zheng K, Wu C, et al. An ensemble method based on selection using bat algorithm for intrusion detection[J]. The Computer Journal, 2018, 61(4): 526-538.

[3] Bhati B S, Chugh G, Al-Turjman F, et al. An improved ensemble based intrusion detection technique using XGBoost[J]. Transactions on Emerging Tele-Communications Technologies, 2021, 32(6): No. e4076.

[4] Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention[J]. Advances in Neural Information Processing Systems, 2014, 27:1-12.

[5] Ahmad I. Feature selection using particle swarm optimization in intrusion detection[J]. International Journal of Distributed Sensor Networks, 2015, 11(10): No. 806954.

[6] Dickson A, Thomas C. Improved PSO for optimizing the performance of intrusion detection systems[J]. Journal of Intelligent & Fuzzy Systems, 2020, 38(5): 6537-6547.

[7] Aziz M R, Alfoudi A S. Feature selection of the anomaly network intrusion detection based on restoration particle swarm optimization[J]. International Journal of Intelligent Engineering & Systems, 2022, 15(5):592-600.

[8] Wei P, Li Y F, Zhang Z, et al. An optimization method for intrusion detection classification model based on deep belief network[J]. IEEE Access, 2019, 7: 87593-87605.

[9] Panigrahi R, Borah S. A detailed analysis of CICIDS2017 dataset for designing intrusion detection systems[J]. International Journal of Engineering & Technology, 2018, 7(3): 479-482.

[10] Goryunov M N, Matskevich A G, Rybolovlev D A. Synthesis of a machine learning model for detecting computer attacks based on the Cicids2017 dataset[J]. Proceedings of the Institute for System Programming of the RAS, 2020, 32(5): 81-94.

[11] Stiawan D, Idris M Y B, Bamhdi A M, et al. CICIDS-2017 dataset feature analysis with information gain for anomaly detection[J]. IEEE Access, 2020, 8:132911-132921.

[12] Salo F, Injadat M, Nassif A B, et al. Data mining techniques in intrusion detection systems: a systematic literature review[J]. IEEE Access, 2018, 6: 56046-56058.

[13] Turnipseed I P. A new scada dataset for intrusion detection research[D]. Starkville: James Worth Bagley College of Engineering, Mississippi State University, 2015.

[14] Rastogi A K, Narang N, Siddiqui Z A. Imbalanced big data classification: a distributed implementation of smote[C]//Proceedings of the Workshop Program of the 19th International Conference on Distributed Computing and Networking, Varanasi, India, 2018: 1-6.

[15] Myles A J, Feudale R N, Liu Y, et al. An introduction to decision tree modeling[J]. Journal of Chemometrics: a Journal of the Chemometrics Society, 2004, 18(6): 275-285.

[16] Biau G, Scornet E. A random forest guided tour[J]. Test, 2016, 25: 197-227.

[17] Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting(version 0.4-2) [DB/OL]. [2015-12-13]. https://gitee.com/mirrors/xgboost/.

[18] 温博文, 董文瀚, 解武杰, 等. 基于改进网格搜索算法的随机森林参数优化[J]. 计算机工程与应用, 2018, 54(10): 154-157.
Wen Bo-wen, Dong Wen-han, Xie Wu-jie, et al. Parameter optimization method for random forest based on improved grid search algorithm[J]. Computer Engineering and Applications, 2018, 54(10): 154-157.

[19] Pattawaro A, Polprasert C. Anomaly-based network intrusion detection system through feature selection and hybrid machine learning technique[C]//The 16th International Conference on ICT and Knowledge Engineering(ICT&KE), Bangkok, Thailand, 2018: 1-6.

[20] 李红亚, 彭昱忠, 邓楚燕, 等. GA与PSO的混合研究综述[J]. 计算机工程与应用, 2018, 54(2): 20-28.
Li Hong-ya, Peng Yu-zhong, Deng Chu-yan, et al. Review of hybrids of GA and PSO[J]. Computer Engineering and Applications, 2018, 54(2): 20-28.

[21] Mohammed M, Mwambi H, Omolo B, et al. Using stacking ensemble for microarray-based cancer classification[C]//International Conference on Computer, Control, Electrical, and Electronics Engineering, Khartoum, Sudan, 2018: 1-8.

[22] 王辉, 李昌刚. Stacking集成学习方法在销售预测中的应用[J]. 计算机应用与软件, 2020, 37(8): 85-90.
Wang Hui, Li Chang-gang. Application of Stacking integrated learning method in sales forecasting[J]. Computer Applications and Software, 2020, 37(8): 85-90.

[23] 张开放, 苏华友, 窦勇. 一种基于混淆矩阵的多分类任务准确率评估新方法[J]. 计算机工程与科学, 2021, 43(11): 1910-1919.
Zhang Kai-fang, Su Hua-you, Dou Yong. A new multi-classification task accuracy evaluation method based on confusion matrix[J]. Computer Engineering & Science, 2021, 43(11): 1910-1919.

[24] Belarbi O, Khan A, Carnelli P, et al. An intrusion detection system based on deep belief networks[C]//International Conference on Science of Cyber Security, Matsue, Japan, 2022: 377-392.

[25] Yao Y, Su L, Lu Z. DeepGFL: deep feature learning via graph for attack detection on flow-based network traffic[C]//IEEE Military Communications Conference(MILCOM), Los Angeles, USA, 2018: 579-584.

[26] Roopak M, Tian G Y, Chambers J. Deep learning models for cyber security in IoT networks[C]//IEEE The 9th Annual Computing and Communication Workshop and Conference, Las Vegas, USA, 2019: 452-457.