第49卷第7期 中国科学技术大学学报 Vol. 49, No. 7
2019年7月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jul. 2019

文章编号:0253-2778(2019)07-0544-11

基于在线自适应极限学习机选择性集成的网络入侵检测

何捷舟[1],刘金平[1,2],张五霞[1],肖文辉[3],唐朝晖[3],徐鹏飞[1]

(1.湖南师范大学智能计算与语言信息处理省重点实验室,湖南长沙 410081;
2.湖南师范大学计算与随机数学教育部重点实验室,湖南长沙 410081;
3.中南大学信息科学与工程学院,湖南长沙 410081)

摘要:随着互联网的普及和网络连接设备与访问方式的多样化,网络入侵方式与手段日趋多样化且变异速度快,传统入侵检测方法在有效性、自适应性和实时性方面难以应对日益复杂网络环境的安全监控要求,为此提出一种基于在线自适应极限学习机(online adaption extreme learning machine, OAELM)选择性学习的网络入侵检测方法(SEoOAELM-NID).首先,提出一种能自动设定最优隐含节点个数且具有在线增量学习功能的OAELM构建方法,采用Bagging策略快速训练出多个具有一定独立性的OAELM子学习器;然后,基于边缘距离最小化原则(margin distance minimization,MDM)对OAELM子学习器的集成增益进行计算;通过选择增益度高的部分OAELM进行选择性集成,获得泛化能力强、效率高的选择性集成学习器用于入侵检测.由于SEoOAELM-NID能自动设定ELM子学习器最优隐节点个数且能根据网络环境变化实现检测模型在线顺序更新,因而能有效适应各种复杂网络环境的入侵检测要求;选择部分最优的子学习器进行集成,保证了最终检测结果的准确性和实效性,同时利用在线数据不断更新检测器.在NSL-KDD数据集上的测试结果表明,相比基于单个学习器以及传统集成学习的网络入侵检测方法,SEoOAELM-NID无论对已知入侵类型还是未知入侵类型均能获得更高的检测率,且识别速度快.

关键词:网络入侵检测;集成学习;在线自适应极限学习机

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2019.07.004

引用格式:何捷舟,刘金平,张五霞,等.基于在线自适应极限学习机选择性集成的网络入侵检测[J].中国科学技术大学学报,2019,49(4):544-554.

HE Jiezhou, LIU Jinping, ZHANG Wuxia, et al. Selective ensemble of online sequential adaption ELMS-based adaptive network intrusion detection[J]. Journal of University of Science and Technology of China, 2019,49(4):544-554.

Selective ensemble of online sequential adaption ELMS-based adaptive network intrusion detection

HE Jiezhou[1], LIU Jinping[1,2], ZHANG Wuxia[1], MA Tianyu[1], TANG Zhaohui[3], XU Pengfei[1]

((1. *Hunan Provincial Key Laboratory of Intelligent Computing and Language Processing, Changsha* 410081, *China*;
2. *Key Laboratory of Computing and Stochastic Mathematics (Ministry of Education), Changsha* 410081, *China*;
3. *School of Information Science and Engineering, Central South University, Changsha* 410083, *China*)

收稿日期:2018-09-21;**修回日期**:2018-12-04

基金项目:国家自然科学基金(61501183, 61771492,61472134),国家自然科学基金-广东联合基金重点(U1701261)、湖南省自然科学基金(2018JJ3349),湖南省研究生科研创新项目(CX2018B312)资助.

作者简介:何捷舟,男,1994年生,硕士生,研究方向:机器视觉与模式识别.E-mail:hdc@smail.hunnu.edu.cn

通讯作者:刘金平,博士/副教授.E-mail:ljp202518@163.com

Abstract: The popularity of the Internet and network equipment and the diversity of access methods have brought great about convenience as well as huge security challenges. The ways and means of network intrusion are becoming more diversified and faster. Traditional intrusion detection methods are unable to meet the security monitoring requirements of an increasingly complex network environment in terms of effectiveness, adaptability and real-time. This paper proposes a network intrusion detection method based on selective learning of the online sequential Adaption Extreme Learning Machines (OAELMs), termed SEoOAELM-NID. Firstly, an OAELM construction method with online incremental update function is proposed, which can automatically set the optimal number of hidden nodes. Bagging strategy is used to train several OAELM sub-learners with certain independence. Then, based on the Margin Distance Minimization (Margin Distance Minimization) guidelines, the OAELM sublearner is integrated into the gain measure, and ensembled by selecting a partial sublearner with high gain. To get a highly Selective Ensemble of OAELM high generalization ability. SEoOAELM-NID has the advantages of automatic optimal setting of hidden nodes and online sequential update of ELM sub-learners, so it can effectively adapt to the intrusion detection requirements of various complex network environments; and by selecting some optimal sub-learners for integration, the accuracy and effectiveness of the final detection results are guaranteed, and online application is used. The test results on the NSL-KDD data set show that SEoOAELM-NID can achieve higher detection rates and fast recognition speeds for known and unknown intrusion types than single learner and traditional ensemble learning-based network intrusion detection methods.

Key words: network intrusion detection; ensemble learning; online adaption ELM

0　引言

随着互联网的普及，网络连接设备和访问方式面临着更多的安全威胁. 2017 年，严重的网络安全事故频现报端. 比如，WannaCry 勒索病毒事件席卷全球、雅虎 30 亿用户账号信息泄露、Facebook 5000 万用户信息泄露等，众多网络安全事件让人们对网络安全的关注达到前所未有的高度. 网络安全甚至事关国家安全与发展. 习近平总书记在 2016 年的网络安全和信息化工作座谈会就曾深刻指出“网络安全是事关国家安全和国家发展、事关广大人民群众工作生活的重大战略问题”.

网络入侵检测(network intrusion detection, NID)作为一种主动的安全防护技术，已经成为网络安全必不可少的一部分. NID 的目标是识别系统内部人员和外部访问者未经授权的使用、误用和滥用计算机系统. 由于计算机系统的连接性越来越强，使得入侵者能够更容易地入侵外部网络，网络连接特征复杂多变，新型的入侵方法层出不穷，使得入侵检测面临更大的挑战.

传统的 NID 采用误用检测，将网络连接数据与建立好的入侵库相匹配，从而发现异常连接. 这种检测技术虽然正确率较高，但无法检测出新型或者变种入侵. 基于异常的入侵检测通过分析正常连接的连接特征，如连接时长，占用资源等，获得正常连接模型；入侵检测时，通过检测网络连接数据与正常连接模型的偏差进行异常报警. 虽然该方法需要一定时间开销进行(正常)网络连接模型学习，但是因其对新型入侵类型有较好的检测效果，近年来广受关注[1-2].

常用的异常检测方法为基于单(学习)模型的检测方法，包括，神经网络、模糊识别等[3-5]. 文献[5]基于稀疏逻辑回归技术提出一种基于特征选择的入侵检测与攻击类型识别方法，通过稀疏正则化优化，并从原始特征变量中选择一个特征子集，用于网络连接数据建模与入侵类型自动分类. 该方法将特征选择和分类组合在统一框架，对已知类型的攻击具有较高的检测准确率，但对于网络环境复杂多变、攻击方式层出不穷的复杂网络安全监测来说，其泛化能力较差. 文献[6]将多种改进的 SVM 算法用于异常检测. 研究表明，虽然 SVM 对于未知的异常网络连接有着较好的识别效果，但其入侵检测时间长，难以达到实时检测的效果.

单模型的异常检测方法因入侵检测模型复杂度

不同各有优劣. 相对来说,复杂检测模型能获得较高识别率,但相应的训练时间较长;简单的入侵检测模型(学习算法)对连接数据特征维数高、入侵方式复杂的网络入侵检测,难以取得有效的检测效果. 近年来,基于多模型的(集成学习)的入侵检测方法受到越来越广泛的关注. 比如,张玲等[7]提出一种基于粗糙集和人工免疫的集成算法,取得了较好的检测效果.

集成学习算法通过集成多个子学习器来提升整体算法的泛化能力. 理论上来说,对于未知网络入侵类型,集成方法远优于基于单个学习器的入侵检测方法[8]. 在实际的应用中,并不是每一个子学习器对于最终的集成学习器都是有利的,那些性能较差的子学习器在训练和测试中会占用大量的时间与空间资源. 如果能选择出其中性能良好的部分子学习器进行有选择性集成,理论上将能获得更好的 NID 性能.

针对网络连接特征复杂多变,各种攻击方式层出不穷,现有的网络入侵检测技术存在的检测速度慢、新型入侵检测率低等问题,本文提出一种基于在线自适应极限学习机(OAELM)选择性集成的网络入侵检测方法,命名为 SEoOAELM-NID. 首先,基于 ELM 快速学习的特点,提出一种具有自动优化设定 ELM 隐含层节点数和增量式在线顺序更新的自适应 ELM(OAELM);然后,基于 Bagging 策略独立训练出多个 OAELM 子学习器;通过细分每一个子学习器对集成算法的增益度对部分 OAELM 子学习器进行选择性集成,从而加强整体算法的泛化能力. 在 NSL-KDD 数据上进行验证,结果表明,本文方法不但能有效检测已知的网络入侵类型,且对未知入侵类型也有较高的识别率.

1 基于 OAELM 选择性集成的 NID

ELM 算法具有学习速度快、泛化能力强等优势近年来受到研究者们的广泛关注[9],基于 Bagging 机制进行 ELM 子学习器集成,可以有效提升 NID 算法的正确率以及识别效率. 直接基于传统 ELM 进行选择性集成以实现 NID 存在如下两大难题.

(Ⅰ)ELM 隐含层节点的优化设定问题. 传统的 ELM 需要依靠人工经验手动设置隐含层节点数目,因而无法保证隐层节点数目设置的有效性;或者需要通过大量的试验来确定,难以保证检测模型对各种数据的适应能力和检测的时效性.

(Ⅱ)NID 检测模型的实时响应问题. 针对网络攻击复杂多变的特性,一个良好的 NID 模型需要能及时响应网络环境的变化,并能根据网络的变化实现模型在线更新.

为解决上述两大难题,本文提出一种具有在线更新和隐层节点自适应能力的 OAELM 子学习器. OAELM 采用自适应 ELM 学习器来自动调节隐含层神经元个数,同时引入在线更新机制提升检测算法的实用性.

本文提出的 SEoOAELM-NID 主要包括基于 Bagging 策略的自适应 ELM 子学习器离线学习,基于 MDM 的子分类器选择性集成,ELM 子学习器在线更新等步骤. SEoOAELM-NID 结构如图 1 所示.

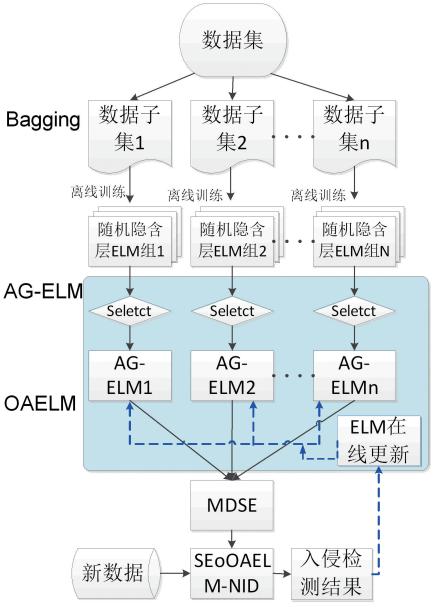

图 1 SEoOAELM-NID 结构

Fig. 1 SEoOAELM-NID structure

1.1 集成学习

近年来,在机器学习领域,集成学习方法受到了广泛的关注. 其通过将具有一定差异性的多个弱学习器采用一定规则进行集成来对最终结果进行共同预测,通过牺牲空间复杂度与时间复杂度来达到提升最终判别正确率的效果,其具有代表性的结合方法 Bagging 和 Boosting 已经被证明是非常有效的[10-12]. 其中 Boosting[13]通过串行化的方式生成子学习器,每一个子学习器都在前一个子学习器的基础上针对识别错误的样例来提升识别效果,该方法强调每一个子学习器之间的依赖关系,并通过串行序列化生成. 代表算法有 Adaboost[14]、XGBOOST[15]、GBDT. Bagging 即 Boostrap Aggregation,其中 Boostrap 是一

种有放回的抽样方法，其抽样策略是简单的随机抽样，所以 Bagging 是一种基于数据随机重抽样的分类器构建方法. 它提倡每一个子学习器都应该尽可能的相互独立，采用可同时生成的分布式并发计算方法进行建立. Bagging 学习策略如图 2 所示.

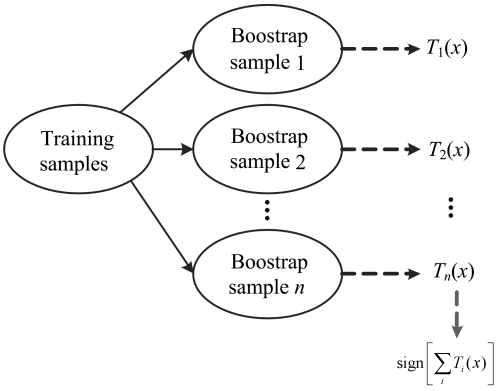

图 2　Bagging 学习策略

Fig. 2　Bagging learning strategy

近年来还出现了许多新的集成算法以及在基础集成算法上的改进算法. 比如，文献[16]基于平均 1-依赖贝叶斯分类器(AODE)算法，提出了平均 1-依赖决策树集成算法(AODT). 通过使用每个输入属性和类别属性共同建立集成学习中的个体决策树分类器. 该方法具有较好的分类性能，具有一定的抗噪性能，但该方法采用各个子分类器的平均值作为最终结果，忽略了每个子学习器在分类性能上的强弱性.

本文充分考虑了各子分学习器性能的差异，提出一种基于 MDM 的选择性集成算法(margin distance minimization selective ensemble, MDMSE)，通过对总体学习器进行集成增益排序选择性能表现良好的部分子学习器作为最终的入侵检测器，有效降低了弱学习器对最终集成结果的不利影响，可以获得稳定的入侵检测结果.

1.2　自适应 OAELM 子学习器

ELM 是一种简单易用、有效的通用单隐层前馈神经网络(single hidden layer feed-forward neural networks，SLFNs)算法，2004 年由南洋理工大学黄广斌教授提出[17]. 其通过随机化生成隐藏层神经元节点参数来计算输出权重矩阵从而解决了传统的前馈神经网络学习算法[18]需要不断迭代修改隐含层参数的问题，提升了运算的速度且具有泛化性能好的优点. 典型的 ELM 模型如图 3 所示.

ELM 具有训练效率高，泛化能力强等优点，克服了传统神经网络算法学习效率低，参数设定繁琐的问题. 相比梯度下降算法，ELM 的神经网络单隐含层需要更多的节点，节点的增加将带来计算和存储的开销. 同时调节隐含层神经元个数也是个繁琐的过程. 为此许多学者在原有 ELM 算法基础上提出了动态 ELM 学习算法，如 AG-ELM(adaption generater ELM)[19]，通过在训练的过程中不断随机生成不同隐藏层节点数目的 ELM，最终选择学习效果最好且节点数目相对较少的 ELM 作为最终的学习器.

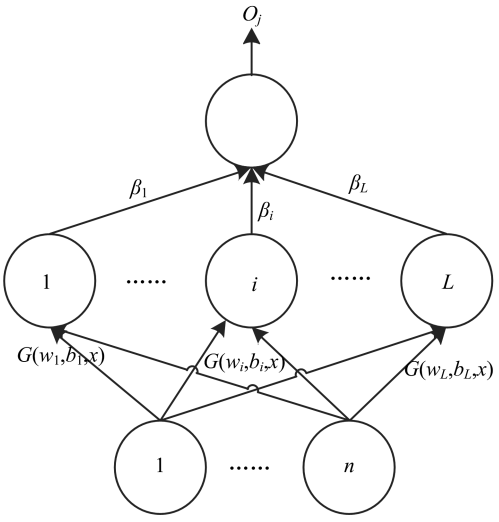

图 3　ELM 网络结构

Fig. 3　ELM network structure

在传统的学习问题中，数据都是预先采集好的，但在网络入侵检测中，入侵方式更新快，NID 系统需要满足对实时的在线数据进行训练并更新，因此本文在原有的 AG-ELM 算法上引入在线序列 ELM(online sequential ELM，OS-ELM)[20]的思想，提出了一种在线自适应极限学习器(online adaption ELM, OAELM).

OAELM 首先采用 AG-ELM 训练出合适的 ELM 网络并计算输出权重. 在随后的在线检测过程中，每当新的数据块被接受，就会重新运行一次 AG-ELM 并得到新的输出权重. 最后新旧输出权重会进行组合从而完成对入侵检测模型的在线更新.

1.2.1　基本的 ELM

对于一个有 L 个隐层节点的 SLFNs，假设有 N 个任意的样本 (X_i, T_i)，其中 $X_i = \{X_{i1}, X_{i2}, \cdots, X_{in}\}^{\mathrm{T}} \in R^n$，$t_i = [t_{i1}, t_{i2}, \cdots, t_{in}]^{\mathrm{T}} \in R^m$，可以表示为

$$\sum_{i=1}^{L} \beta_i g(W_i \cdot X_j + b_i) = O_j, j = 1, \cdots, N \quad (1)$$

式中，$g(*)$为特征映射函数，β_i 为输出权重，$W_i=[w_{i1},w_{i2},\cdots w_{in}]^{\mathrm{T}}$ 为输入权重，b_i 为第 i 个隐层单元的偏置. $W_i\cdot X_j$ 表示 W_i 和 X_j 的内积.

为使输出结果与真实标签误差最小，其损失函数可以表示为

$$\sum_{j=1}^{n}\|O_j-t_j\|=0 \tag{2}$$

即存在输出权重 β_i，输入权重 W_i 和偏置 b_i 使得

$$\sum_{i=1}^{L}\beta_i g(W_i\cdot X_j+b_i)=t_j,j=1,\cdots,N \tag{3}$$

可以矩阵表示为

$$H\beta=T \tag{4}$$

式中，H 是隐层节点的输出，β 是输出权重，T 是期望输出.

$$H(W_1,\cdots W_L,b_1,\cdots b_L,X_1,\cdots X_L)=\begin{pmatrix} g(W_1\cdot X_1+b_1) & \cdots & g(WL\cdot X_1+bL)\\ \vdots & \cdots & \vdots\\ g(W_1\cdot X_N+b_1) & \cdots & g(W_L\cdot X_N+b_L)\end{pmatrix}_{N\times L} \tag{5}$$

通过求解，可以获得输出权重 β

$$\hat{\beta}=H^{\dagger}T \tag{6}$$

式中，$H^{\dagger}$ 是矩阵 H 的 Moore-Penrose 广义逆. 研究表明，所获得的 $\hat{\beta}$ 是范数最小的唯一解.

1.2.2 自适应极限学习机(AG-ELM)

ELM 作为一种高效的 SLFNs 网络学习方法，虽然不需要迭代修改神经元参数，但对单隐层神经元数目的设置也至关重要，过多的神经元会造成大量的计算开销，并可能产生过拟合，而少量的神经元又很难达到理想的泛化效果.

自适应极限学习机(adaption generater ELM，AG-ELM)[19]很好地解决了这一问题.

AG-ELM 采用自适应的方式自动增减隐含层神经元个数，通过不断的迭代选择更合适的神经元个数的网络作为最终的选择结果.

定义 1.1 伪均匀分布. 令 $\theta=(w,b,\beta)$ 是一个 χ 上的随机数，θ 的概率密度为 $P(x)$. 如果在 χ 上的每一个 x 都存在 $P(x)\neq 0$，那么可以称随机变量 θ 为伪均匀分布. 其中 $\chi=R^d\times R\times R$，$d>2$. w 为 ELM 隐含层的输入权重矩阵，b 为偏置矩阵，β 为输出权重矩阵.

基于伪均匀分布定义，AG-ELM 迭代步骤如下：

Step 1 基于伪均匀分布 χ 随机产生一组参数 $\theta_1^{(1)}=(w_1^{(1)},b_1^{(1)},\beta_1^{(2)})$ 且令

$$\varphi_1^{(1)}=\beta_1^{(1)}g(w_1^{(1)}x+b_1^{(1)}).$$

Step 2 基于伪均匀分布 χ^2 随机产生两组随机参数 $(\theta_1^{(2)},\theta_2^{(2)})=((w_1^{(2)},b_1^{(2)},\beta_1^{(2)}),(w_2^{(2)},b_2^{(2)},\beta_2^{(2)}))$，且令

$$\widetilde{\varphi}(2)_1=\beta_1^{(2)}g(w_1^{(2)}x+b_1^{(2)}) \tag{7}$$

$$\widetilde{\varphi}(2)_2=\beta_1^{(2)}g(w_1^{(2)}x+b_1^{(2)})+\beta_2^{(2)}g(w_2^{(2)}x+b_2^{(2)}) \tag{8}$$

记

$$\varphi_1^{(2)}=\{\varphi_1^{(1)},\text{if}\ \|\varphi_1^{(1)}-y\|_p\leqslant\|\widetilde{\varphi}(2)_1-y\|_p\ \widetilde{\varphi}_1^{(2)},\text{otherwise} \tag{9}$$

$$\varphi_2^{(2)}=\{\varphi_1^{(2)},\text{if}\ \|\varphi_1^{(2)}-y\|_p\leqslant\|\widetilde{\varphi}(2)_2-y\|_p\ \widetilde{\varphi}_2^{(2)},\text{otherwise} \tag{10}$$

式中，y 为标签值，即将第一步产生单神经元 $\theta_1^{(1)}$ 的网络与第二步产生单神经元 $\theta_1^{(2)}$ 的网络比较，将误差小的记作 $\varphi_1^{(2)}$，然后将 $\varphi_1^{(2)}$ 与第二步所产生含 $\theta_1^{(2)}\theta_2^{(2)}$ 两个神经元的网络比较，将误差较小值记作 $\varphi_2^{(2)}$.

Step s $(s=3,4,5\cdots,n,\cdots)$：基于伪均匀分布 χ^s 随机产生 s 组随机参数 $(\theta_1^{(s)},\theta_2^{(s)},\cdots,\theta_s^{(s)})$，令

$$\widetilde{\varphi}(s)_1=\beta_1^{(s)}g(w_1^{(s)}x+b_1^{(s)}) \tag{11}$$

$$\widetilde{\varphi}(s)_2=\sum_{i=1}^{2}\beta_i^{(s)}g(w_i^{(s)}x+b_i^{(s)}) \tag{12}$$

$$\vdots$$

$$\widetilde{\varphi}(s)_s=\sum_{i=1}^{s}\beta_i^{(s)}g(w_i^{(s)}x+b_i^{(s)}) \tag{13}$$

记

$$\varphi_1^{(s)}=\{\varphi_{s-1}^{(s-1)},\text{if}\ \|\varphi_{s-1}^{(s-1)}-y\|_p\leqslant\|\widetilde{\varphi}(s)_1-y\|_p\ \widetilde{\varphi}_1^{(s)},\text{otherwise} \tag{14}$$

$$\varphi_s^{(s)}=\{\varphi_{s-1}^{(s)},\text{if}\ \|\varphi_{s-1}^{(s)}-y\|_p\leqslant\|\widetilde{\varphi}(s)_s-y\|_p\ \widetilde{\varphi}_s^{(s)},\text{otherwise} \tag{15}$$

最后能够得到一个函数序列 $\{\varphi_1^{(1)},\varphi_1^{(2)},\varphi_2^{(2)},\cdots,\varphi_1^{(s)},\varphi_2^{(s)},\cdots,\varphi_l^{(m)},\cdots\}$.

记 $\Phi_n=\varphi_l^m$，产生网络的总数记为 n，则 $n=\frac{1}{2}\cdot m(m-1)+l$，记得到的网络序列为 $\Phi_n(x)$，则对任意 n 都满足：$\|\Phi_n-y\|\leqslant\|\Phi_{n-1}-y\|$. 则由 AG-ELM 算法产生的网络序列 $\Phi_n(x)$ 具有很好的泛化效果.

与其他结构增长性或递减性 ELM 算法不同，AG-ELM 从最简单的单个神经元结构开始自适应调节，并且所有参数随机产生，当产生更优的新网络时，会采用新网络替代原有网络. AG-ELM 无需人

为调节任何参数，其隐藏层神经元为自适应增长.

1.2.3　在线顺序极限学习机

考虑到网络入侵方式多样，入侵类型变换快，入侵检测算法的动态更新至关重要，为此引入在线顺序极限学习机(online sequential extreme learning machine, OSELM)[20]思想，使其具有增量式极限学习机的泛化效率. 相比一般的 ELM，OSELM 可以针对新的数据进行批量在线更新，从而保证检测模型能根据网络环境的变化进行自适应更新. OSELM 的算法流程描述如下：

Step 1　初始化阶段：取 $k=0$，k 表示用于 ELM 训练的一个数据段(一个用于 ELM 训练的数据子集). 给定激活函数 $g(x)$和隐含层节点数 L，选择训练样本 $D_0=\{(x,y)\}$初始化网络，求得初始输出权值，即

$$\beta_0=K_0^{-1}H_0^{\mathrm{T}}T_0 \tag{16}$$

式中，K_0 为 $H_0^{\mathrm{T}}H_0$.

Step 2　在线学习阶段：基于最新获得的第 $k+1$ 个数据段对 ELM 模型进行增量式更新. $\beta^{(k+1)}$ 满足

$$\beta^{(k+1)}=\underset{\beta}{\operatorname{argmin}}\left\{\left\|\begin{bmatrix}H_k\\H_{k+1}\end{bmatrix}\beta-\begin{bmatrix}T_k\\T_{k+1}\end{bmatrix}\right\|^2\right\} \tag{17}$$

得

$$\beta^{(k+1)}=K_{k+1}^{-1}\begin{bmatrix}H_{\mathrm{k}}\\H_{k+1}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}T_k\\T_{k+1}\end{bmatrix}=$$
$$K_{k+1}^{-1}(K_{k+1}\beta^{(k)}-H_{k+1}^{\mathrm{T}}H_{k+1}\beta^{(k)}+H_{k+1}^{\mathrm{T}}T_{k+1})=$$
$$\beta^{(k)}+K_{k+1}^{-1}H_{k+1}^{\mathrm{T}}(T_{k+1}-H_{k+1}\beta^{(k)}) \tag{18}$$

式中，$K_k=\begin{bmatrix}H_k\\H_{k+1}\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}H_k\\H_{k+1}\end{bmatrix}$，$K_{k+1}=K_k+H_{k+1}^{\mathrm{T}}H_{k+1}$.

Step 3　$k=k+1$，继续收集新的训练数据集，返回 Step 2 在线学习阶段，更新输出权值.

1.3　基于边缘距离最小化的选择性集成

为了获得良好的集成效果，很多集成算法采用了大量的子学习器. 由于过多的子学习器并不一定具有更好的集成效果，反而会增大时间和存储的损耗. 周志华等基于遗传算法提出了“选择性集成学习”的概念[21]，将采用随机抽样生成的训练子集生成的弱学习器 t 进行权值编码 w_i，组成权值向量 $w=\{w_1,w_2,\cdots,w_t\}$. 最后采用遗传算法进行求解. 遗传算法基于随机变异进行求解，有一定的概率落入局部最优解，同时庞大的运算量也会影响算法的实时性.

本文基于文献[22]提到的边缘距离最小化(margin distance minimization)原理提出了一种新的选择性集成学习算法 MDMSE.

MDMSE 可以计算出每个子学习器对整体算法性能提升的增益度量. 从而选择增益度高的子学习器进行选择性集成. 从以下几个定义着手，对 MDMSE 思想进行说明.

定义 1.2　边界距离最小化(MDM). 给定包含 N 条数据的样本集 D，定义分类器 t 的特征向量 C^t 是其在数据集 D 上判别的正确度，其第 i 部分为

$$C_i^{\mathrm{t}}=2\oplus(h_t(x_i)==y_i)-1,\ (x_i,y_i)\in Z \tag{19}$$

当分类器在第 i 个例子上的判别正确时(即判别结果 $h_t(x_i)$ 与标签 y_i 相等)则 C^t 为 1 否则为 -1.

在集成学习中，所有子学习器的平均特征向量可以定义为

$$\overline{C}=\frac{1}{T}\sum_{t=1}^{T}C^{\mathrm{t}} \tag{20}$$

定义平均特征向量 $\overline{C}$ 的第 i 个元素是第 i 个例子的边界. 其表达了不同子学习器在该例子上分类正确与不正确的差异性，正则化在$[-1,1]$上. 对于多分类，第 i 部分的边界值可以表示为 $(1-\mathrm{edge}(i))$，其中 $\mathrm{edge}(i)$ 是分类正确的类与分类错误类的差异性，其正规化范围为$[0,1]$[23]，因此当特征向量在 N 维空间的第一象限时，其在样本集 D 上的分类结果完全正确.

定义 1.3　N 维空间优化点 O. 为使平均向量位于第一象限，可以在第一象限中的任意位置选择一个点 O，该点在任意维上具有相同的位移 P，即

$$O_i=p\ \text{with}\ i=1,\cdots,N\ \text{and}\ 0<p<1 \tag{21}$$

本文的目的就是将最终集成的选择器平均特征向量尽量靠近点 O，则每次选择的分类器可以表示为

$$S_u=\underset{k}{\operatorname{argmind}}(O,\frac{1}{T}(C^k+\sum_{t=1}^{u-1}C^t)),\ k\in E_T\backslash S_{u-1} \tag{22}$$

式中，$d(v,u)$是点 v 与点 u 的欧几里得距离. S_u 是第 U 次选择的子学习器对减少特征向量 $\overline{C}$ 到目标点 O 的距离增益度.

P 为常量且足够小(例如 $P\sim0.075$)，如果 P 值过大，则每一次选择新的子学习器进行集成时都会考虑所有样本的增益，从而很难提升算法的有效性. 反之足够小的 P 值可以使简单的例子快速分类正确，且不会对后面的选择造成影响；从而保证集成的子学习器具有一定的差异性与泛化性. 在本文实

验，P 设为 0.75 以保证算法的有效性.

本文提出的 SEoOAEL-NID 基于 OAELM 可在线更新且具有快速学习和检测的优点，采用 Bagging 机制训练多个 OAELM，同时采用 MDMSE 算法对集成子学习器进行部分选择性集成，以减小弱学习器对最终集成结果的影响，同时保证集成算法的最大泛化能力，基于 OAELM 的选择性集成算法流程如算法 1.1 所示：

算法 1.1 基于 OA-ELM 的选择性集成算法

将训练集按 bagging 机制进行随机抽样，形成 N+1 个子集.

采用 N 个子集训练出 N 个 AG-ELM 子学习器，使其隐藏层神经元自适应调节.

利用第 N+1 个子训练集采用 MDMSE 对 N 个子学习器进行选择性学习形成集成学习器 ST

在线部署 ST，在识别的同时对新数据进行批量学习采用 OSELM 的学习机制线更新 OA-ELM 子学习器

2 实验结果及分析

2.1 数据集介绍及评价指标

本文采用 NSL-KDD[24] 数据集进行实验验证，NSL-KDD 是 KDD99[25] 数据集的改进版本. 数据集中包含了 4 大类 39 小类的异常入侵链接，其中 22 种异常链接包含在训练集中，另外 17 种在测试集中，用于对本文算法泛化性能的检测.

数据集包含了 41 维特征和一个标签项，且存在字符型特征. 本实验首先采用预处理将字符型特征转换为数值型，然后对数据集进行标准化归一化. 本文实验的软硬件环境配置如表 1，所用数据集分布如表 2 所示.

表 1 实验环境配置

Tab. 1 Configuration of experimental environment

硬件配置	Intel Core i5-3230 CPU @2.60GHz 8.0GB RAM
软件环境	Ubuntu16.04 64 位操作系统 Python3.0+pyCharm

表 2 NSL-KDD 分布

Tab. 2 NSL-KDD distribution

	Train Set	Test Set
DOS	45 927	7 456
Probe	11 656	2 421
R2L	995	2 756
U2R	52	200
Normal	67 343	9 711

对于入侵检测算法性能评判，常用正确率来评价分类算法的好坏. 正确率确实是很直观的评价指标，但一个算法正确率高并不能代表算法就一定好. 对于入侵检测问题，异常的连接在庞大的网络连接数据中只占很小一部分. 特别在实际的网络安全监测中，可能一万条正常的连接才会出现一个异常入侵连接. 对于一个不加思考的分类器，将所有的连接都视为好的，那么其正确率将达到 99.99%. 当真的有异常入侵时，这个分类器并不能察觉到，所以对于这种现实数据不均衡的实验，本文采用多个评判指标从各方面对算法的好坏进行评估. 常用模型评价术语以及混淆矩阵如表 3 所示.

表 3 混淆矩阵表

Tab. 3 Confusion matrix table

		预测情况	
		Yes	No
实际情况	Yes	TP	FN
	No	FP	TN

表 3 中 TP(true positives)表示被正确地划分为正列的个数；FP(false positives)表示被错误地划分为正例的个数；FN(false negatives)表示被错误地划分为负例的个数；TN(true negatives)表示被正确地划分为负例的个数.

常用的评价指标有：正确率(accuracy)、错误率(error rate)、精准率(precition)、召回率(recall)以及 F_1 值.

accuracy=(TP+TN)/(TP+FP+TN+FN)；

error rate=(FP+FN)/(TP+FP+TN+FN)；

precision=TP/(TP+FP)；

recall=TP/(TP+FN)；

F_1=2 precision * recall/(precision+ recall).

考虑到网络安全的重要性以及第二次人工检测. 我们应该最大化减小异常连接的漏报率，因此本文在采用正确率与 F_1 值来评估算法之外，定义了一个新的评价标准漏报率，即

$$\text{miss rate} = 1\text{-TN}/(\text{TN}+\text{FP}).$$

式中，正确率用来直接衡量检测器对网络入侵检测的优劣. F_1 值则可以有效地衡量检测器的稳定性，其同时考虑精准率和召回率，当其值越大检测器越稳定. 漏报率是指所有异常的数据中有多少没有被分类器检测出来，用来衡量检测器的泛化性.

2.2 实验流程

实验采用 SEoOAELM-NID 算法在 NSL-KDD 数据集上进行实验，主要流程如下：

(Ⅰ)输入：DATA={(X_1, y_1),(X_2, y_2),…,

$(X_m, y_m)\}$.

(Ⅱ)对于数据集进行有放回的行和列随机采样,行采样 m 次($m=M$. 为输入样本集样本个数)分为 T 份,列采样 N 次(N 为数据集特征数).

(Ⅲ)for $t=1,2\cdots T-1$:

使用采样集 $TDt(i=1,..,T)$进行 OAELM 训练得到子学习器 $Gt(x)$

(Ⅳ)for $u=1,2,\cdots,U$(部分集成子学习器个数):

for $t=1,2,\cdots,T-1$:

进行 MDMSE 估计公式如下:

$$S_u=\underset{k}{\operatorname{argmin}}(O,\frac{1}{T}(C^k+\sum_{t=1}^{u-1}C^t))$$

每次将增益度最大的 OAELM 子学习器加入选择子学习器集合 ST.

(Ⅴ)ST 为最终训练出来的集成学习器.

(Ⅵ)在线部署 ST 检测器,当接收到新的数据块时,重新运行 OAELM 得到新的输出权重,然后使用新旧权重进行组合更新 OAELM.

2.3　实验结果

实验主要包含两个部分:①验证性实验:针对 SEoOAELM-NID 算法与其他算法在数据集 NSL-KDD 上的分类正确率、F_1 值和漏报率进行了比较验证;②参数调节实验:对算法中的参数进行了调节验证,以期进一步提升算法性能.

2.3.1　验证性结果

通过调用 Python[26] 的机器学习框架中的算法在 NSL-KDD 数据集上进行训练,将其检测结果与 SEoOAELM-NID 算法进行实验对比,其准确率、漏报率以及 F_1 值如表 4(其中个别算法所用数据集特征采用了 PCA 特征降维来提高其准确率与减少训练时长),其中 SVM 使用的核函数是径向基函数,而本文算法中隐含神经元激活函数采用 sigmoid 函数. 为避免学习算法不稳定造成误差,对以下算法重复实验了 30 次,结果为其平均值.

通过表 4、5 可以看出,随机森林与梯度决策树作为集成算法确实能获得比单学习器更高的正确率,但同时也大大增加了时间的开销;而泛化性能强、漏报率低的 SVM 算法又很难保证算法的正确率. 本文所提出的 SEoOAELM-NID 采用计算速度快泛化能力强的 OAELM 作为子学习器且采用 Bagging 的分布式计算,在提高正确率的同时保证了较低的漏报率以及检测时长.

表 4　SEoOAELM-NID 与其他算法正确率比较

Tab. 4　SEoOAELM-NID comparing with the correctness rate of other algorithms

算法	评估指标	分类类别					平均
		PROBING 入侵连接	DOS 入侵连接	R2L 入侵连接	U2L 入侵连接	正常连接	
SEoOAELM-NID	正确率	0.971	0.979	0.987	0.966	0.972	0.975
	漏报率	0.130	0.163	0.124	0.175	0.143	0.147
	F1	0.79	0.81	0.77	0.80	0.79	0.79
随机森林[27]	正确率	0.935	0.940	0.947	0.932	0.938	0.938
	漏报率	0.586	0.614	0.587	0.623	0.619	0.606
	F1	0.62	0.55	0.58	0.61	0.59	0.59
朴素贝叶斯[28]	正确率	0.768	0.792	0.773	0.783	0.785	0.780
	漏报率	0.415	0.408	0.399	0.414	0.378	0.403
	F1	0.55	0.55	0.62	0.58	0.58	0.58
梯度决策提升树[29]	正确率	0.923	0.927	0.927	0.931	0.922	0.926
	漏报率	0.674	0.576	0.585	0.633	0.609	0.615
	F1	0.55	0.57	0.61	0.58	0.58	0.58
支持向量机[30]	正确率	0.912	0.918	0.922	0.916	0.923	0.918
	漏报率	0.243	0.222	0.195	0.194	0.213	0.213
	F1	0.71	0.73	0.76	0.77	0.71	0.74
C-ELM[31]	正确率	0.933	0.936	0.931	0.937	0.941	0.934
	漏报率	0.373	0.380	0.371	0.376	0.377	0.375
	F1	0.68	0.64	0.66	0.66	0.65	0.66

表 5 SEoOAELM-NID 与其它检测算法在运行时间上的比较

Tab. 5 SEoOAELM-NID comparing with the speed of other algorithms

	Train time	Test time
SEoOAELM-NID	0.592	0.0243
随机森林	3.589	0.261
朴素贝叶斯	5.812	0.531
梯度决策提升树	22.623	0.165
支持向量机	120.343	4.675

通过 F_1 值可以发现 SEoOAELM-NID 拥有更好的稳定性，因此说明其在检测尽量多的异常连接的同时能够保证检测的准确性.

在 NSL-KDD 测试集中包含了 17 类未训练的异常网络连接，本实验对比了不同算法在未知类型的异常链接上检测的效果，其实验结果如表 6 所示. 由表 6 可知，本文所提出 SEoOAELM-NID 算法在未知类型的异常连接上仍具有较低的漏报率，且将 17 种未知类型的异常连接均检测出来，具有很好的泛化性能.

表 6 未知入侵连接检测数据

Tab. 6 Unknown intrusion connection detection data

算法	检测到的类型数	未知类型检测漏报率
SEoOAELM-NID	17	0.14
随机森林	14	0.28
朴素贝叶斯	8	0.63
梯度决策提升树	15	0.24
支持向量机	13	0.31

2.3.2 参数设置对算法性能的影响

本文基于 OAELM 提出的 SEoOAELM-NID 算法，其隐藏层神经元个数自适应调节，但仍有一些参数需要手动设置，如 Bagging 集成中对列的随机抽样 feature 抽样率，选择性集成中子学习器个数 U 等，而 SEoOAELM-NID 算法中各参数的合理设置也十分重要. 为此本文对以上两种参数的设置进行了实验分析.

(Ⅰ)feature 抽样率

Bagging 集成算法的 feature 抽样率设置对算法正确率与漏报率影响如表 7 所示.

表 7 features 抽样率

Tab. 7

feature	N	$0.5N$	$0.01N$	$\mathrm{Log2}(N)$
正确率	0.973 8	0.972 2	0.974 6	0.974 4
漏报率	0.142	0.139	0.144	0.141

因本文所采用的 OAELM 算法具有很强的泛化性且 NSL-KDD 数据集数量庞大特征维度较高，所以表 7 所示 features 抽样率对实验结果的影响并不大.

(Ⅱ)选择性学习器个数 U

对于 SEoOAELM-NID 算法，初始化 Bagging 集成为 100 个子学习器. 在 MDMSE 算法部分，通过调节选择性学习器个数 U 对算法性能进行参数验证如图 4 所示.

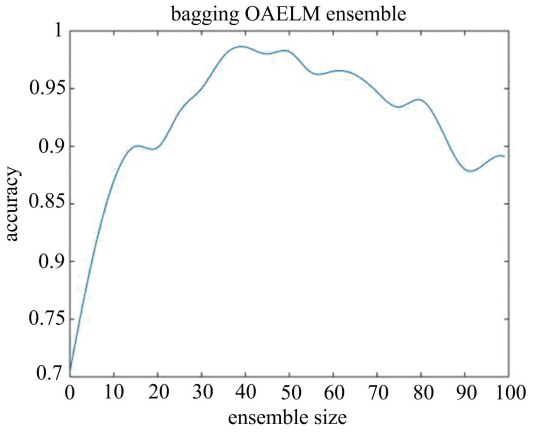

图 4 Bagging OAELM 集成

Fig. 4 Bagging OAELM ensemble

由图 4 可知，并不是集成的子学习器数目越多越好，过多的弱学习器的加入反而会影响最终的集成效果. 本文实验中，当选择性集成的子学习器个数为 40 时，算法的准确率达到最高；当子学习器个数继续增加，其算法准确率反而降低. 从而证明了选择性学习相对于完全集成学习算法的优势所在.

3 结论

本文基于具有在线更新且隐含节点自适应的极限学习机 OAELM，采用 MDMSE 原则对其进行选择性学习，提出了一种网络入侵检测算法 SEoOAELM-NID. 本文所提方法同时具有 OAELM 的快速学习且泛化能力强的特性以及 Bagging 集成算法的分布式并发训练的效果，训练时间短检测速度快，具有良好的实时检测效果. 在 NSL-KDD 数据集上实验结果表明，该算法具有较高的正确率与较低的漏报率. 其选择性集成算法对于网络入侵检测在保证识别正确率的情况下大大缩短了训练与识别的时间.

参考文献(References)

[1] HAMAMOTO A H, SAMPAIO L D H, ABR O T, et al. Network anomaly detection system using genetic algorithm and fuzzy logic[J]. Expert Systems with Applications, 2018, 92(1): 309-402.

[2] JAVAID A, NIYAZ Q, SUN W Q, et al. A deep learning approach for network intrusion detection system[C]// Proceedings of 9th EAI International Conference on Bio-Inspired Information and Communications Technologies. Brussels, Belgium: ACM, 2016: 21-26.

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

[4] SULTANA N, CHILAMKURTI N, PENG W, et al. Survey on SDN based network intrusion detection system using machine learning approaches[J]. Peer-to-Peer Networking and Applications, 2018, 1(1-2): 1-9.

[5] SHAH R A, QIAN Y, KUMAR D, et al. Network intrusion detection through discriminative feature selection by using sparse logistic regression[J]. Future Internet, 2017, 9(4): 81-96.

[6] THASEEN I S, KUMAR C A. Intrusion detection model using fusion of chi-square feature selection and multi class SVM[J]. Journal of King Saud University-Computer and Information Sciences, 2016, 29(4): 462-472.

[7] 张玲,白中英,罗守山, et al. 基于粗糙集和人工免疫的集成入侵检测模型[J]. 通信学报, 2013, 1(9): 166-176.

[8] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 171-173.

[9] 林宇鹏, 谢智歌, 徐凯, 等. 基于超限学习机的快速癌症检测方法[J]. 中国科学技术大学学报, 2018, 48(2): 154-160.
LIN Yupeng, XIE Zhige, XU Kai, et al. Fast cancer diagnosis based on extreme learning machine[J]. Journal of University of Science and Technology of China, 2018, 48(2): 154-160.

[10] ANGELO P, DRUMMOND A C. A survey of random forest based methods for intrusion detection systems [J]. ACM Computing Surveys, 2018, 51(3): 3641.

[11] 文泽波, 康宇, 曹洋, 等. 基于随机森林特征选择的视频烟雾检测[J]. 中国科学技术大学学报, 2017, 47(8): 653-664.
WEN Zebo, KANG Yu, CAO Yang, et al. Features selection for video smoke detection using random forest [J]. Journal of University of Science and Technology of China, 2017, 47(8): 653-664.

[12] 孙艳歌, 王志海, 原继东, 等. 基于信息熵的数据流自适应集成分类算法[J]. 中国科学技术大学学报, 2017, 47(7): 575-582.
SUN Yange, WANG Zhihai, YUAN Jidong, et al. Adaptive ensemble classification algorithm for data streams based on information entropy[J]. Journal of University of Science and Technology of China, 2017, 47(7): 575-582.

[13] SABZEVARI M, SU REZ A. Vote-boosting ensembles[J]. Pattern Recognition, 2018, 83(11): 119-133.

[14] YANG Z, XU L, CAI Z, et al. Re-scale AdaBoost for attack detection in collaborative filtering recommender systems [J]. Knowledge-Based Systems, 2016, 100(1): 74-88.

[15] CHEN T Q, GUESTRIN C. XGBoost: A scalable tree boosting system [C]// Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining. San Francisco, USA: ACM, 2016: 785-794.

[16] 周传华, 王清, 吴科主, 等. 平均 1-依赖决策树集成算法[J]. 电子学报, 2010, 38(2): 434-438.

[17] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: A new learning scheme of feedforward neural networks[C]// International Joint Conference on Neural Networks. Budapest, Hungary: IEEE, 2004: 985-990.

[18] PENG K X, YANG J B, TUO X G, et al. Research on PGNAA adaptive analysis method with BP neural network[J]. Modern Physics Letters B, 2016, 30(32-33): 1650386.

[19] ZHANG R, LAN Y, HUANG G B, et al. Universal approximation of extreme learning machine with adaptive growth of hidden nodes [J]. IEEE Transactions on Neural Networks & Learning Systems, 2012, 23(2): 365-371.

[20] LIMA A R, HSIEH W W, CANNON A J. Variable complexity online sequential extreme learning machine, with applications to streamflow prediction[J]. Journal of Hydrology, 2017, 555(1): 983-994.

[21] ZHOU Z H, WU J X, TANG W. Ensembling neural networks: Many could be better than all[J]. Artificial Intelligence, 2002, 137(1-2): 239-263.

[22] MARTINEZMUOZ G, HERNANDEZLOBATO D, SUAREZ A. An analysis of ensemble pruning techniques based on ordered aggregation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2009, 31(2): 245-259.

[23] MARTíNEZ-MUñOZ G, SUáREZ A. Aggregation ordering in bagging[C]// Proceedings of the IASTED International Conference on Artificial Intelligence and Applications. Innsbruck, Austria: ACTA, 2004: 258-263.

[24] REVATHI S, MALATHI A. A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection[J]. International Journal of Engineering Research and Technology, 2013, 12(2): 1848-1813.

[25] TAVALLAEE M, BAGHERI E, LU W, et al. A detailed analysis of the KDD CUP 99 data set[C]// Proceedings of the 2nd International Conference on Computational Intelligence for Security & Defense Applications. Ottawa, Canada: IEEE, 2009: 53-58.

[26] KRAMER O. Scikit-Learn[M]. Springer International Publishing, 2016: 45-53.

[27] CHANG Y, LI W, YANG Z M. Network intrusion detection based on random forest and support vector machine [C]// International Conference on Computational Science and Engineering. Guangzhou, China: IEEE, 2017: 635-638.

[28] 孙栓柱, 宋蓓, 李春岩, 等. 一种基于贝叶斯后验的异常值在线检测及置信度评估算法[J]. 中国科学技术大学学报, 2017, 47(8): 644-652.
SUN Shuanzhu, SONG Bei, LI Chunyan, et al. An online outlier detection and confidence estimation algorithm based on Bayesian posterior ratio[J]. Journal of University of Science and Technology of China, 2017, 47(8): 644-652.

[29] LI L, YU Y, BAI S, et al. Towards effective network intrusion detection: A hybrid model integrating Gini index and GBDT with PSO[J]. Journal of Sensors, 2018, 2018(6): 1-9.

[30] TENG S, WU N, ZHU H, et al. SVM-DT-based adaptive and collaborative intrusion detection[J]. Journal of Automatica Sinica, 2018, 5(1): 108-118.

[31] WANG C R, XU R F, LEE S J, et al. Network Intrusion Detection Using Equality Constrained-Optimization-Based Extreme Learning Machines[J]. Knowledge-Based Systems, 2018, 147(1): 68-80.