计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900101

基于特征重要度二次筛选的 DDoS 攻击随机森林检测方法

李娜娜[1]　王　勇[1]　周　林[1]　邹春明[2]　田英杰[3]　郭乃网[3]
1 上海电力大学计算机科学与技术学院　上海 200090
2 公安部第三研究所　上海 200031
3 国网上海市电力公司电力科学研究院　上海 200120
(764529188@qq.com)

摘　要　特征选择是攻击检测算法中的一种重要方法,该方法多采用交叉验证递归特征消除(Recursive Feature Elimination with Cross-Validation,RFECV)技术,并通常结合机器学习算法使用。但该算法多用于选取单模型特征,其性能也极易受特征量、学习器的变化而波动,因其计算量大,该算法的分类准确率也仍需提高。针对上述问题,文中提出了一种基于特征重要度二次筛选的 DDoS 攻击随机森林检测方法。首先,该算法对原始数据集进行预处理并提取特征;其次,该算法为了从所选模型中选择最相关的变量,使用 RF 变量重要度准则,利用随机森林的重要性评分对变量进行排序;然后,在随机森林特征排序的基础上,对变量计算累积重要性并得到最重要变量;接着,使用所筛选出的最重要变量再次进行训练以生成分类模型,从而得出一组新的重要变量并将其定义为当前变量;最后,通过重要度准则,计算累积重要性来得到最终的最佳变量,从而有效地去除异常点,避免局部最优,进而实现对 DDOS 攻击的精准分类检测。实验结果表明,该方法具有较高的准确度和精确度,能够实现对正常流量以及各种 DDoS 攻击流量的精准分类,适用于在大数据下检测 DDoS 攻击。

关键词:DDoS 攻击检测;特征提取;重要度准则;机器学习;随机森林
中图法分类号　TP309.2

DDoS Attack Random Forest Detection Method Based on Secondary Screening of Feature Importance

LI Na-na[1],WANG Yong[1],ZHOU Lin[1],ZOU Chun-ming[2],TIAN Ying-jie[3] and GUO Nai-wang[3]
1 College of Computer Science and Technology,Shanghai University of Electric Power,Shanghai 200090,China
2 The Third Research Institute of Ministry of Public Security,Shanghai 200031,China
3 Institute of Electric Power Research,State Grid Shanghai Electric Power Company,Shanghai 200120,China

Abstract　Feature selection is an important method for attack detection algorithms. This method mostly uses cross-validation recursive feature elimination (Recursive Feature Elimination with Cross-Validation,RFECV) technology,and is usually combined with machine learning algorithms. However,this algorithm is mostly used to select single-model features,and its performance is also very susceptible to fluctuations due to changes in feature quantities and learners. Due to the large amount of calculation,the classification accuracy of this algorithm still needs to be improved. In response to the above problems,this paper proposes a random forest detection method for DDoS attacks based on the secondary screening of feature importance. Firstly,the algorithm preprocesses the original data set and extracts features. Secondly,in order to select the most relevant variables from the selected model,the algorithm uses the RF variable importance criterion and the random forest importance score to rank the variables. Then,on the basis of random forest feature ranking,the cumulative importance of the variables is calculated and the most important variables are obtained. Then,the most important variables selected are used for training again to generate a classification model,and a new set of important variables is defined as the current variable. Finally,the final optimal variable is obtained through the importance criterion and the cumulative importance again,which effectively removes the abnormal points and avoids the local optimum,thereby realizing accurate classification and detection of DDOS attacks. Experimental results show that this method has high accuracy and precision,can accurately classify normal traffic and various DDoS attack traffic,and is suitable for

基金项目:国家自然科学基金面上项目(61772327);上海自然科学基金面上项目(20ZR1455900);上海市科委科技创新行动计划(18511105700);上海市科委电力人工智能工程技术研究中心项目(19DZ2252800);奇安信大数据协同安全国家工程实验室开放课题(QAX-201803);浙江大学工业控制技术国家重点实验室开放式基金(ICT1800380)
This work was supported by the General Project of National Natural Science Foundation of China(61772327),General Project of Shanghai Natural Science Foundation of China (20ZR1455900),Shanghai Science and Technology Commission Science and Technology Innovation Action Plan (18511105700),Shanghai Science and Technology Commission Power Artificial Intelligence Engineering Technology Research Center Project (19DZ2252800),Qi'anxin Big Data Collaborative Security National Engineering Laboratory Open Project(QAX-201803) and Open-end Fund of State Key Laboratory of Industrial Control Technology,Zhejiang University(ICT1800380).
通信作者:王勇(wy616@126.com)

detecting DDoS attacks under big data.

Keywords DDoS attack detection，Feature extraction，Importance criterion，Machine learning，Random forests

1 引言

分布式拒绝服务(DDoS)攻击是一种网络攻击，此类攻击具有破坏性强、涉及面广、实施方便、难以追踪和防范等特点，且与其他网络攻击不同，DDoS 攻击只需要大量的僵尸和少量的网络安全知识就可以发起有效的攻击[1]。因此，及时、正确地识别 DDoS 攻击迫在眉睫。

通常来讲，入侵检测系统主要分为两类[2]：基于主机的入侵检测和基于网络的入侵检测。无论采取哪种方法，都需要提取和筛选主机或网络在正常工况和受攻击情况下的特征，因此特征的提取和选择指异常检测的关键环节。特征提取一般通过数学方法将数据从高维特征空间映射到低维特征空间，典型的方法有主成分分析(Principal Component Analysis，PCA)、线性判别分析 (Linear Discriminant Analysis，LDA)[3]等机器学习算法。特征选择指使用某种评价准则从原始特征空间中选出特征子集，是一种数据预处理方式。理想的特征选择算法需要实现去除无关、弱相关和冗余，并保留弱相关非冗余和强相关特征[4]。

本文提出基于特征重要度二次筛选的 DDoS 攻击随机森林检测方法，该方法对提取的网络流量统计特征并计算其重要性，选择出最佳变量，然后将此算法用于检测 DDoS 流量。实验结果表明，该方法具有较高的检测精度(PREC)。

2 问题分析

文献[5]提出了一种基于协议分析和聚类的分布式拒绝服务攻击检测模型。该模型采用数据挖掘算法对协议信息元素进行分析。该攻击检测方法的优点是不需要人工构建数据，同时保持了较高的检测率。然而，对于大型网络来说，单位时间内的网络连接数量为 1 万到 100 万不等。

文献[6]基于集成学习中的 Stacking 学习法思想，提出了一种 CA-LR 学习法。在对原始训练数据集进行 K 折交叉验证的前提下，将传统随机森林(Random Forest，RF)算法中固有的决策树模型与逻辑回归模型结合成新模型，虽然改进随机森林算法的分类模型在预测准确率上比基于传统随机森林算法的分类模型更优，但使用了 10 个随机森林，72.875%的准确率依旧有待提升。

文献[7]提出了一种基于组合相关度的随机森林 DDoS 攻击检测方法。根据攻击流的非对称性和半交互性定义网络流组合相关度(Combination Correlation Degree，CCD)，该相关度以地址相关统计(Address Correlation Statistics，ACS)特征以及单向流半交互度(Unidirectional Flow Semi Interaction，UFSI) 二元组来描述网络流的特点。然后，提出基于 CCD 特征序列的遗传算法对 RF 中决策树的最大数量和最大深度两个关键参数进行优化，再对参数优化的 RF 模型进行训练，以生成分类模型用于检测攻击。虽然此分类模型的准确率保持在 90%以上，误报率保持在 0.4%以下，但漏报率却在 15%左右。

3 基于变量重要度的机器学习特征选择算法

随机森林是一种有监督的学习算法，它建立大量的随机决策树并将它们合并在一起进行预测。每棵树都训练有一个随机的标记样本集的子集。在分类过程中，模型中所有树中投票最多的类代表分类器的结果[8]。随机森林也是一种集成学习算法[9]，它将决策树作为基分类器，创建 Bagging 集成，并将随机属性选择机制加入到决策树训练中。随机森林算法将自身的变量重要性度量作为对高维数据的特征选择工具，其既具有良好的鲁棒性又具备较高的学习效率，能使异常数据或缺失值数据得到很好地处理，因此近年来这一算法被广泛应用于各种分类、预测、特征选择以及异常点检测问题中[10]。因此本文中的特征选择采用了随机森林的算法机制设计合理的特征选择算法，利用其变量重要度准则进行特征自动选择，算法流程图如图 1 所示。

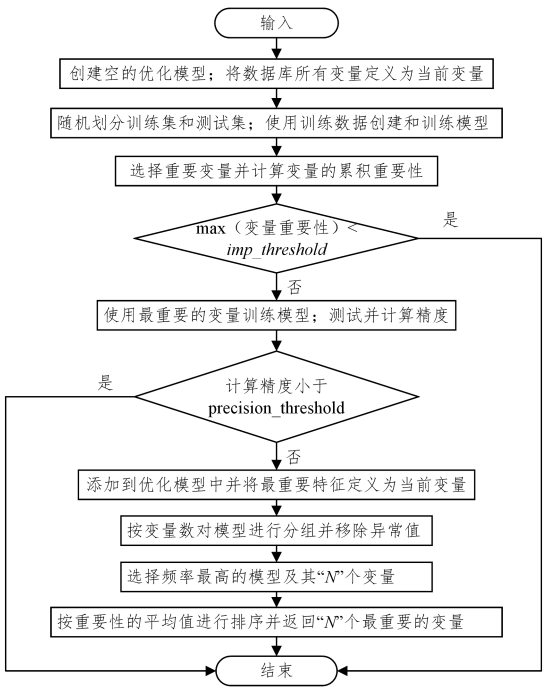

图 1 特征选择算法的流程图

Fig. 1 Flow chart of feature selection algorithm

3.1 相关数据集

KDD 99(加州大学欧文分校 1998－99)和 LBNL(劳伦斯伯克利国家实验室和 ICSI 2004－2005)数据集已经年代久远。本文采用 ISC-XIDS2012 数据集中具有正常活动行为的数据流以及在实验室环境中使用诸如 Hping3，Hulk，Goldeneye 和 Slow httptest 等工具获得的具有 DoS 行为的数据流量组成的数据集[11]，因为它们包含较多现代攻击手段和 DoS 技术，并且在这些数据集中，有一个较高容量的正常流量和各种类型的攻击，包括隐形应用层攻击。在这种更真实的情况下，系统虽然会出现一些检测故障，但仍能取得较好的性能。

3.2 特征提取

在监督分类技术中，需要一组训练分类器模型的数据集，这个集合通常被定义为签名数据集。数据集的每个实例都有一组特性或标签相关联的变量，目前广泛使用的网络流量采样协议(如 NetFlow[12]和 sFlow)在采样过程中仅使用了一部分变量，其通常使用的部分变量包括：源端口、目标端口、传输

层协议、IP 数据包大小和 TCP 标志。同时，主要使用这 5 个变量又导出了 33 个变量[11]，共同构成了在线流量数据集的特征，如表 1 所列。

表 1 变量提取

Table 1 Variable extraction

	Variable	Detail
1	*Ip_proto*	Normalized protocol number
2	*ip_len_mean*	Mean of IP length
3	*ip_len_median*	Median of IP length
4	*ip_len_var*	Variance of IP length
5	*ip_len_std*	Stand. deviation of IP length
6	*ip_len_entropy*	Entropy of IP length
7	*ip_len_cv*	Coeff. of variation of IP length
8	*ip_len_cvq*	Quantile coeff. of IP length
9	*ip_len_rte*	Rate change of IP length
10	*sport_mean*	Mean of src port
11	*sport_median*	Median of src port
12	*sport_var*	Variance of src port
13	*sport_std*	Stand. deviation of src port
14	*sport_entropy*	Entropy of src port
15	*sport_cv*	Coeff. of variation of src port
16	*sport_cvq*	Quantile coeff. of src port
17	*sport_rte*	Rate change of src port
18	*dport_mean*	Mean of dest. Port
19	*dport_median*	Median of dest. Port
20	*dport_var*	Variance of dest. Port
21	*dport_std*	Stand. deviation of dest. Port
22	*dport_entropy*	Entropy of dest. Port
23	*dport_cv*	Coeff. of variation of dest. Port
24	*dport_cvq*	Quantile coeff. of dest. port
25	*dport_rte*	Rate change of dest. Port
26	*tcp_flags_mean*	Mean of TCP flags
27	*tcp_flags_median*	Median of TCP flags
28	*tcp_flags_var*	Variance of TCP flags
29	*tcp_flags_std*	Stand. deviation of TCP flags
30	*tcp_flags_entropy*	Entropy of TCP flags
32	*tcp_flags_cv*	Coeff. of variation of TCP flags
32	*tcp_flags_cvq*	Quantile coeff. of TCP flags
33	tcp_flags_rte	*Rate change of TCP flags*

3.3 实验及分析

在基于变量重要度的机器学习特征选择方法中，算法的参数设置如下：1 000 轮，99% 的可变重要性，95% 的全局精度，85% 的每类精度（$ne=1\ 000$，$i=0.99$，$p=0.95$，$pc=0.85$）。

实验结果如图 2 所示，根据横坐标中模型数量的比较可以看出，模型数量达到最多 4 123 时，纵坐标对应的使用变量个数为 19，即 RF 算法大多数被测试的模型使用了 19 个特征变量。为了从所选模型中选择最相关的变量，使用基于变量重要度的机器学习特征选择算法的最终结果，如图 3 所示，按照特征重要度大于 0.001 的范围，最终选择了 15 个特征。

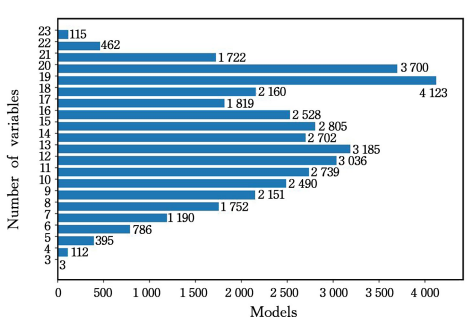

图 2 变量数量与模型数量

Fig. 2 Models Vs. number of variables

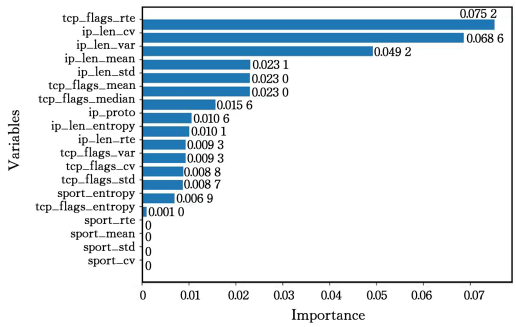

图 3 变量选择

Fig. 3 Variable selection based

3.4 评估指标

本文使用文献[13-14]中的精度（PREC）、召回率（REC）和 F-度量（F1）指标来评估系统性能。PREC 为测量时避免假阳性的能力，而 REC 为测量系统灵敏度。F1 是 PREC 和 REC 之间的调和平均值。1）真阳性（TP）是正确预测的攻击流量；2）真阴性（TN）是正确预测的正常流量；3）假阳性（FP）是错误预测的正常流量；4）假阴性（FN）是错误预测的攻击流量。

指标的计算式如式（1）—式（3）所示：

$$PREC=\frac{TP}{TP+FP} \tag{1}$$

$$REC=\frac{TP}{TP+FN} \tag{2}$$

$$F1=2\times\frac{PREC\times REC}{PREC+REC} \tag{3}$$

基于随机森林算法的模型评价指标如图 4 所示。

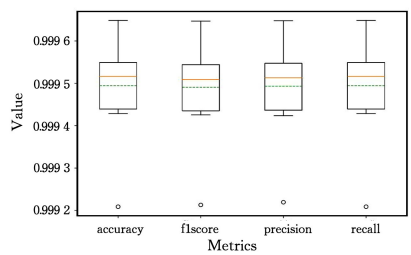

图 4 模型评价指标（电子版为彩色）

Fig. 4 Model evaluation index

实验模型评价结果表明，RF 算法的准确度、精确度均达到了 99.95%，采用 KDD 99 数据集进行对比实验，由于此数据集与本文算法的特征不同，因此致使精度参数的最优设置仅达到 $p=0.5$，$pc=0.5$，且结果中准确度、精确度也只达到 85% 左右。由此可知，本文算法中的特征提取部分对算法的结果具有至关重要的作用。然而，由图 4 可看出，本文算法的准确度、精确度虽高，但算法对异常点（见图 4 中圆点）较为敏感，每个指标的训练和测试结果间的误差也较大（红线、绿线分别表示训练值和测试值）。

4 基于特征重要度二次筛选的 DDoS 检测算法

第 3 节所述的特征算法的实验结果对异常点较为敏感，且训练与测试值相差较大。针对上述问题，在基于随机森林的特征选择算法的基础上进行了特征选择优化，实现在一次重要特征提取的基础上进行二次筛选。即从训练模型中选择最重要的特征，并计算出累积重要性（cumulative_impor-

tances)，根据所设阈值 i(imp_threshold)来判定是否提取特征，并以此作为当前变量再次训练模型，进而二次通过重要度准则来计算累积重要性，从而选择出最终的最佳特征，实现有效去除异常点。算法输入为：$ne=1\,000$，$i=0.99$，$p=0.95$，$pc=0.85$。

算法 1 基于特征重要度的二次筛选算法

输入：(d,i,p,pc,ne,t)
输出：N 个重要变量
1. 创建空的优化模型
2. for i←1 to 设定轮数 do
3. 将数据库所有特征定义为当前变量
4. while 训练、测试集变量所属当前变量 do
5. 分割数据集
6. 使用训练数据创建和训练模型
7. 利用 RF 重要度准则对特征排序
8. 计算训练模型中特征的累积重要性
9. If max(cumulative_importances)＜imp_threshold, then Exit loop;
10. end
11. 使用最重要的变量训练、测试模型并计算准确度
12. 利用 RF 重要度准则再次对特征排序
13. 计算累积重要性
14. If max(cumulative_importances)＜imp_threshold, then Exit loop;
15. end
16. 找出 cumulative_importances＞＝imp_threshold 的特征
17. 提取最重要特征
18. if precision＜precision_threshold then Exit loop;
19. end
20. 将当前模型添加到优化模型集
21. end while
22. end
23. 按特征数对模型进行分组
24. 从分组模型集中移除异常值
25. 选择频率最高的模型组及其特征数“N”
26. 根据步骤 13 中计算的重要性的平均值对变量进行排序
27. 返回“N”个最重要的变量

5 实验及分析

实验结果表明，标签分类混淆矩阵的分类精确度达到了 99.96%，具体如图 5 所示。

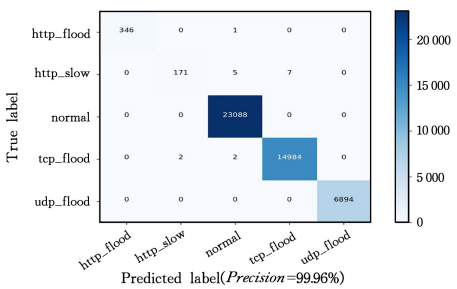

图 5 标签分类混淆矩阵

Fig. 5 Label classification confusion matrix

在模型数量达到最多 4020 时，其对应的使用变量个数为 20，相比第 3 部分 RF 算法仅增加了一个使用变量，具体如图 6 所示。二次特征筛选的最终结果如图 7 所示，与图 3 所示的 RF 一次特征选择结果相比，重要度大于 0.01 的多数重要特征在二次筛选之后均提升至 0.05 以上，同时最佳变量也从 15 个增加到 20 个。

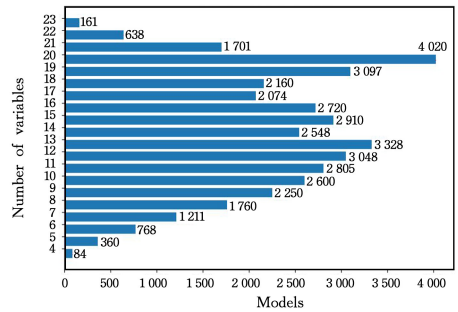

图 6 变量数量与模型数量

Fig. 6 Models Vs. number of variables

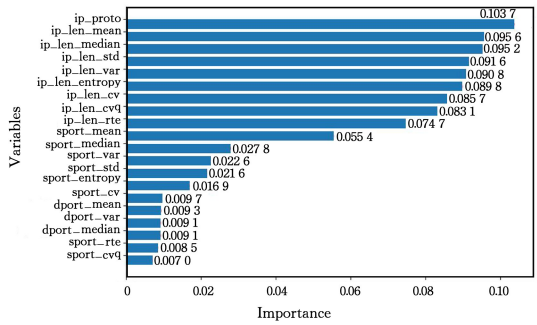

图 7 RF 特征重要度二次筛选算法变量选择

Fig. 7 RF feature importance secondary filtering algorithm variable selection

如图 8 和表 2 所示，进行二次特征筛选的算法在特征数量保持不变的情况下，消除了异常点的影响，每个特征的重要度显著提升，且各项指标的训练、测试结果也均有明显提升，其误差也更小，从而验证了模型具有更高的分类检测能力。

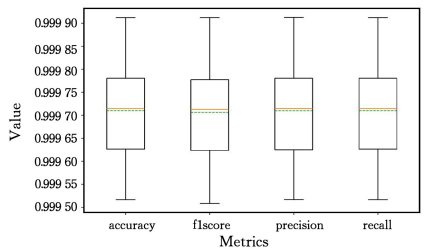

图 8 RF 特征重要度二次筛选算法模型评价指标

Fig. 8 Evaluation index of RF feature importance secondary screening algorithm model

表 2 算法结果对比

Table 2 Comparison of algorithm results

Algorithm	Best models	Bestvariables	Accuracy/%	Precision/%
RF	19	15	99.95	99.95
改进 RF	20	20	99.97	99.97

结束语 本文提出的基于特征重要度二次筛选的 DDoS 攻击随机森林检测方法，通过对特征选择算法的改进来实现特征的重要性评级、选择最佳特征数量的优化，并生成 DDoS 攻击分类模型来检测攻击。实验结果表明，该方法与文献中现有的一些相关方法相比，能更好地提取攻击流量特性及其

(下转第 476 页)

Delegation[C]// Requirements Engineering Conference (RE). 2005:167-176.
[18] GIORGINI P, MASSACCI F, MYLOPOULOUS J, et al. Requirements Engineering meets Trust Management: Model, Methodology, and Reasoning[C]// Proc. of iTrust'04, LNCS 2995. Springer-Verlag, 2004:176-190.
[19] TIDWELL T, LARSON R, FITCH K, et al. Modeling Internet Attacks[C]// Proceedings of the 2001 IEEE Workshop on Information Assurance and Security United States Military Academy. 2001.
[20] LAMSWEERDE A V. Elaborating security requirements by construction of intentional anti-models[C]// ICSE. 2004:148-157.
[21] SHOSTACK A. Threat Modeling: Designing for Security [M]. John Wiley & Sons, 2014.
[22] SCANDARIATO R, WUYTS K, JOOSEN W. A descriptive study of microsofts threat modeling technique[J]. Requirements Engineering, 20(2):163-180.
[23] Information technology-Security techniques-Information security riskmanagement[M]. ISO, 2011.
[24] KREBS B. Email Attack on Vendor Set Up Breach at Target [EB/OL]. http://krebsonsecurity.com/2014/02/email-attack-on-vendor-set-up-breach-at-target/.

DU Jin-lian, born in 1972, Ph.D, associate professor, is a member of China Computer Federation. Her main research interests include software reliability and security requirement, text data analysis, and physical based simulation.

JIN Xue-yun, born in 1972, lecturer. Her main research interest includes software automation.

(上接第 467 页)

引起的正常流状态异常变化的特征，并能有效地避免局部最优和去除异常点的不利影响。提出的检测方法比同类检测方法具有更高的准确度和精确度，从而实现了对正常流量以及攻击流量的精准分类。

参 考 文 献

[1] WANG C, ZHENG J, LI X Y. Research on DDoS attacks detection based on RDF-SVM[C]// International Conference on Intelligent Computation Technology and Automation (ICICTA). 2017:161-165.
[2] ZHANG W A, HONG Z, ZHU J W, et al. A survey of network intrusion detection methods for industrial control systems[J]. Control and Decision, 2019, 34(11):2277-2288.
[3] XU J Z, WU Z H, XU Y, et al. Face recognition combining PCA, LDA and SVM algorithms[J]. Computer Engineering and Applications, 2019, 55(18):34-37.
[4] LI Z Q, DU J Q, NIE B, et al. Summary of Feature Selection Methods[J]. Computer Engineering and Applications, 2019, 55(24):10-19.
[5] GAO N, FENG F D, XIANG J. A data-mining based dos detection technique[J]. Jisuanji Xuebao(Chinese Journal of Computers), 2006, 29(6):944-951.
[6] PEI J T. DDOS Attack Detection based on machine learning and Big Data Real-time Computing analysis [D]. Beijing: Beijing University of Technology, 2019.
[7] LI M Y, TANG X Y, CHENG J R, et al. Random forest DDoS attack detection method based on combination correlation [J]. Journal of Zhengzhou University (Science Edition), 2019, 51(2):23-28.
[8] SYLVESTER E, BENTZEN P, BRADBURY I R, et al. Applications of random forest feature selection for fine-scale genetic population assignment [J]. Evolutionary Applications, 2018, 11(2):153-165.
[9] ZHAO L, CHEN Z, HU Y, et al. Distributed feature selection for efficient economic big data analysis [J]. IEEE Transactions on Big Data, 2008, 32(2):164-176.
[10] YANG C C, XU X, HUAN J, et al. Feature selection method of student portrait based on random forest[J]. Computer Engineering and Design, 2019, 40(10):2827-2834.
[11] FILHO F, SILVEIRA F, JUNIOR A, et al. Smart Detection: An Online Approach for DoS/DDoS Attack Detection Using Machine Learning[J]. Security and Communication Networks, 2019 (12):1-15.
[12] Cisco systems netflow services export version 9 [EB/OL]. https://www.rfc-editor.org/info/rfc3954.
[13] PARK S H, GOO J M, JO C H. Receiver operating characteristic (ROC) curve: practical review for radiologists[J]. Korean Journal of Radiology, 2004, 5(1):11-18.
[14] MARTIN D, POWERS W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation [J]. Journal of Machine Learning Technologies, 2011, 2(1):37-63.

LI Na-na, born in 1992, postgraduate. Her main research interests include robot safety and Informationsecurity.

WANG Yong, born in 1973, Ph.D, professor. His main research interests include power system virus analysis and defense.