基于流量异常分析多维优化的入侵检测方法

刘新倩[1,3], 单 纯[2,4*], 任家东[1,3], 王 倩[1,3], 郭嘉伟[1,3]

[1]燕山大学信息科学与工程学院 秦皇岛 中国 066001
[2]北京理工大学 北京 中国 100081
[3]河北省软件工程重点实验室 秦皇岛 中国 066001
[4]北京市软件安全工程技术重点实验室 北京 中国 100081

摘要 入侵检测系统在检测和预防各种网络异常行为的过程中, 海量和高维的流量数据使其面临着低准确率和高误报率的问题。本文提出一种基于流量异常分析多维优化的入侵检测方法, 该方法在入侵检测数据的横向维度和纵向维度两个维度进行优化。在横向维度优化中, 对数量较多的类别进行数据抽样, 并采用遗传算法得到每个类别的最佳抽样比例参数, 完成数据的均衡化。在纵向维度优化中, 结合特征与类别的相关分析, 采用递归特征添加算法选择特征, 并提出平均召回率指标评估特征选择效果, 实现训练集的低维高效性。基于优化的入侵检测数据, 进一步通过训练数据集得到随机森林分类器, 在真实数据集 UNSW_NB15 评估和验证本文提出的算法。与其他算法相比, 本文算法具有高准确率和低误报率, 并在攻击类型上取得了有效的召回率。

关键词 入侵检测框架; 多维优化; 数据抽样; 递归特征添加算法; 遗传算法参数优化; 随机森林
中图法分类号 TP393.08 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2019.01.02

An intrusion detection method based on multi-dimensional optimization of traffic anomaly analysis

LIU Xinqian[1,3], SHAN Chun[2,4*], REN Jiadong[1,3], WANG Qian[1,3], GUO Jiawei[1,3]

[1] Department of Information Science and Engineering, Yanshan University, Qinhuangdao 066001, China
[2]Beijing Institute of Technology, Beijing 100081, China
[3] Hebei Key Laboratory of Software Engineering, Qinhuangdao 066001, China
[4]Beijing Key Laboratory of Software Security Engineering Technology, Beijing 100081, China

Abstract In the process of detecting and preventing various network anomaly behaviors, intrusion detection system is facing the problem of low accuracy and high false alarm rate due to the massive and high-dimensional traffic data. An intrusion detection method based on multi-dimensional optimization of traffic anomaly analysis is proposed, in which both horizontal and vertical dimensions of intrusion detection dataset are optimized. In horizontal dimensions optimization, those categories with a large number are sampled and the optimal sampling proportion parameters of each category are obtained by genetic algorithm. Data equalization is accomplished. In vertical dimensions optimization, combining with the correlation analysis of features with label, recursive features addition algorithm is adopted to select features, and the average recall is proposed to evaluate the effect of features selection. The low-dimensional and high-efficient training data set is achieved. Based on optimized intrusion detection dataset, the random forest classifier is obtained by training dataset, and the real data set UNSW_NB15 is used to evaluate and validate the proposed method. Compared with other algorithms, the proposed algorithm has high accuracy and low false alarm rate, and effective recall rate on attack category is obtained.

Key words intrusion detection framework; multi-dimensional optimization; data sampling; recursive features addition; genetic algorithm parameter optimization; random forest

1 引言

网络服务在各个领域中都得到了广泛的应用, 各种新用户、新设备和其他事物不断地连入网络[1]。伴随网络的迅速扩展, 多种异常事件和攻击行为频繁发生, 对网络性能和安全造成了巨大的损害[2]。多

通讯作者: 单纯, 讲师, Email:sherryshan@bit.edu.cn。
本课题得到国家重点研发计划项目(No. 2016YFB0800700)，国家自然科学基金(No. 61472341, No. 61772449, No. 61572420, No. 61807028, No. 61802332)，河北省自然科学基金(No. F2016203330)，博士后科研择优资助项目(No. B2017003005)资助。
收稿日期：2018-09-30；修改日期：2018-11-24；定稿日期：2018-12-14

种安全机制, 例如杀毒软件、防火墙技术、用户身份验证和访问控制[3]等技术被设计并应用到计算机网络中, 用来防御异常行为并检测潜在的风险[4]。然而, 由于网络攻击行为的发生频率和强度不断增强, 现有的安全机制已经不能有效地保护网络空间的安全[5]。基于这种原因, 入侵检测系统作为网络安全的第二道防线, 不断地被研究并广泛应用。

在计算网络中, 入侵检测系统用来监测和分析网络行为, 标识出任何与正常事件的偏差行为。一般情况下, 入侵检测通常分为两类: 基于误用和基于异常的入侵检测系统[6]。基于误用的入侵检测系统通过构建正常行为与攻击行为的规则库, 可以很好地检测已有的攻击行为, 具有良好的准确性和误报率, 但这种检测方法的缺点在于不能有效地检测出新型的攻击行为[7]。基于异常的入侵检测系统根据历史数据中正常网络行为建立检测模型, 尝试去检测正常行为的偏差, 当偏差值超出一定阈值时, 系统定义异常行为发生。这种检测方法能够检测新型和未知的异常行为, 但缺点在于高误报率和低准确率[8]。为了更好的改善入侵检测系统的性能, 多种技术被应用到入侵检测研究中, 例如抽样技术、特征选择技术以及分类算法。

抽样技术从数据层面对整个不均衡数据集进行研究, 解决数据分布不均衡问题, 将提高算法的分类效果[9]。抽样技术主要分为 3 种: 欠采样方法、过采样方法和混合采样方法。欠采样通过减少训练集中多数类的样本个数达到类别间的相对平衡。过采样通过增加训练集中少数类的样本个数, 使类别达到平衡。混合采样是欠采样方法和过采样方法的混合使用。一些研究者已经将抽样技术应用到入侵检测研究中。Hamed 等人[1]采用 SMOTE 技术来处理不平衡的网络数据集得到新的平衡训练数据集, SMOTE 技术是一种过采样技术, 通过一定策略合成少数类样本来达到平衡数据的目的。Wathiq 等人[10-12]提出改进的 K-means 方法来对每个多数类进行欠采样, 选取部分数据来代表整体数据, 所选数据的数据量要远远少于原始数据集。抽样技术的应用在很大程度上改善了异常检测效果, 提高了准确性, 但海量数据的处理需要极高的计算复杂度和时间消耗。并且目前入侵检测领域中应用的抽样方法在抽样数据量的选择没有客观的评价, 无法保证抽样后的数据集具有最优的检测效果。除此之外, 入侵检测数据的均衡化问题研究较少, 因此, 该问题仍然具有较大的研究价值和应用价值。

特征选择技术已成为入侵检测研究中不可或缺的一部分[13]。在实际检测过程中, 特征选择技术能大大降低训练和检测时间, 同时选择的特征子集能提高或至少保持异常检测效果[14-16]。特征选择技术主要分为 2 种: 分类器独立方法(过滤式)和分类器依赖方法(包装式和嵌入式)[17]。过滤式特征选择方法主要采用统计学和信息论方法, 例如主成分分析、信息增益和互信息等[18]。Mohamed 等人[19]指出基于信息论的特征选择方法不关注特征与分类器之间的交互, 导致对信息的重要性评价过高, 因此造成了选择的特征是冗余和不相关的, 为了解决这一问题, 提出了结合互信息和“最大化最小化”方法来选择特征, 减小特征冗余和不相关。Gao 等人[20]提出一种混合方法结合最小化特征冗余和最大化分类信息来选择特征, 并在两者之间取得一个平衡, 既保证了特征与类别之间的相关性, 又保证了特征与分类器之间的交互。Shadi 等人[21]采用信息增益作为评估指标从 NSL-KDD 数据集的 41 个特征中选择 8 个特征作为最终的评估子集。这种特征选择方法的缺陷在于, 会对某些数据值更多的特征的重要性进行过高估计, 选择出冗余和不相关的特征。为了弥补这一缺陷, 一些研究者提出了分类器依赖特征选择方法。该方法采用机器学习算法来评估特征重要性从而选择出最具代表性的特征子集[17]。Chaouki 等人[17]采用逻辑回归作为评估模型, 遗传算法作为搜索策略, 去选择最小且最相关的特征子集。Tarfa 等人[22]提出一种基于支持向量机的递归特征选择方法, 其中支持向量机的成本函数变化量作为特征重要性的评估标准。分类器依赖方法通过与分类器进行交互, 减少了特征冗余, 提高了相关性, 但其计算代价要远远高于过滤式方法。

在构建入侵检测分类器过程中, 多种机器学习算法被应用, 例如模糊逻辑(Fuzzy logic, FL)、贝叶斯网络(Bayesian network, NB)、人工智能神经网络(Artificial neural networks, ANN)、随机森林(Random forest, RF)和支持向量机(Support vector machine, SVM)作为单一的分类器用于异常检测的分类问题中[23-28]。Abdulla 等人[29]采用加权的 WOAR-SVM(Weighted one-against-rest SVM)构建入侵检测分类器, 并与多种 SVM 分类器(OAO-SVM(One-against-one SVM)、DAG-SVM(Directed acyclic graph SVM)等)进行比较, 说明 WOAR-SVM 更能抵消错误, 得到更好的检测结果。Ugo 等人[30]提出采用受限的玻尔兹曼机并结合能量模型来构建正常流量模型, 从而检测网络异常。现在的许多研究也提出混合分类器和多层分类器的观点。Vijayanand 等人[31]采用多

个 SVM 分类器对数据进行多次二分类, 判断该数据是否属于某种已知的行为, 否则, 该数据属于新型攻击行为。Shadi 等人[21]结合 J48、Meta Pagging、Random tree 和 REPTree 等算法构建混合的投票模型, 得到较好的准确率。混合分类器在一定程度上能提高入侵检测的分类效果, 但由于目前对于入侵检测数据集并没有一个统一的收集指标和生成指标, 因此, 混合分类器不易扩展到不同的入侵检测数据中。不同分类器对异常检测效果影响很大, 因此, 构建一个有效的异常检测分类器是有必要的。

为了改善目前入侵检测的低准确率和高误报率问题, 同时针对上述研究中存在的问题, 本文提出一种基于流量异常分析多维优化的入侵检测方法。本文的创新之处在于以统计分析为基础, 从数据的横向和纵向两个维度进行优化, 得到高效的入侵检测数据。数据的横向优化指的是数据条目的抽样, 对数量较多的类别进行随机抽样, 结合遗传算法和准确性指标, 实现数据集类别间的均衡性。数据的纵向优化指的是数据特征的选择, 结合过滤式和嵌入式特征选择方法的优势, 实现数据集的低维性。最终在优化的数据集上, 采用随机森林模型构建异常检测分类器, 该分类器具有良好的异常检测效果。真实的入侵检测数据集 UNSW-NB15 用来评估本文算法的检测效果。

本文的结构组织如下。第二章对入侵检测数据集进行统计分析。第三章阐述了本文提出的方法。第四章分析数据集和实验结果。第五章对本文的工作进行总结。

2 数据统计分析

这一部分将对入侵检测数据进行统计分析。数据的统计分析对数据的深入研究和学习算法的应用具有重要的意义。统计分析主要集中在 2 个方面: (1)数据总体分布情况; (2)特征与类别之间的相关性。本文研究异常检测领域中的新的数据集 UNSW-NB15, 该数据集在 2015 年由澳大利亚安全实验室采集网络流量数据生成的, 比传统的入侵检测数据集 KDD1999 和 NSL-KDD 更符合现代网络流量的特点(协议组成、服务、带宽等)[32], 因此, 本文主要针对入侵检测数据集 UNSW-NB15 进行研究。该数据集包括 42 个特征(3 个符号特征和 39 个数值特征)和 1 个类别标签(正常行为和 9 种攻击行为)[32]。

数据总体分布情况: 采用饼状图来可视化数据的总体分布, 图 1 显示了不同类别数据所占的比例, 其中 Normal 类型所占的比例最大, 攻击类型 Generic 和 Exploits 同样占有较大的比例, 而 Worm、Shellcode、Backdoor 和 Analysis 类型占的比例仅为的 0.1%、0.6%、1.1%和 1.0%。因此, 异常检测数据是一个极其不均衡的数据集。

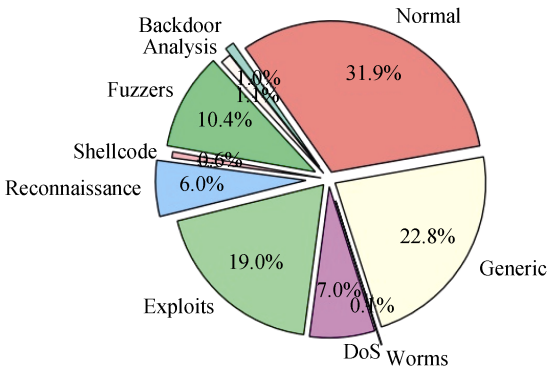


图 1 UNSW_NB15 数据集的整体分布

Figure 1 The overall distribution of UNSW_NB15 data set

特征与类别之间的关联分析: 分析特征与类别之间的相关性, 特征与类别的相关性越大, 越有利于分类效果, 降低特征冗余[15]。本文采用交叉列联检验来描述特征与类别的相关性, 交叉列联又称为交互分类, 是指同时依据两个变量的值, 将所研究的个案分类。交互分类的目的是将两变量分组, 然后比较各组的分布状况, 以寻找变量间的相关关系。并针对不同类型的变量采用不同的相关性指标。符号特征与类别的相关性采用 V 系数评估。V 系数常用于名义变量之间的相关系数计算。V 系数的计算公式是由卡方统计量修改而来的, 其计算公式如(1)所示。

$$V=\sqrt{\frac{\chi^2}{N(K-1)}}, \tag{1}$$

其中, χ^2 表示卡方统计量, N 表示样本个数, K 表示列联表中行数和列数较小的实际数。V 系数越大, 说明名义变量之间相关性越强。数值特征与类别的相关性采用 *Eta* 系数评估。*Eta* 系数常用于名义变量和数值变量的相关性检验。*Eta* 的平方表示组间差异所解释的因变量的方差的比例, 其具体的计算公式如下所示。

$$\begin{aligned} Eta^2 &= \frac{SS_b}{SS_t} \\ &= \frac{SS_t - SS_w}{SS_t} \\ &= \frac{\Sigma(y-\bar{y})^2 - \Sigma(y-\bar{y_i})^2}{\Sigma(y-\bar{y})^2}, \end{aligned} \tag{2}$$

其中，SS_b 表示组件差异，$SS_t=\sum(y-\bar{y})^2$ 表示总体差异，$SS_w=\sum(y-\overline{y_i})^2$ 表示组内差异，$\bar{y}$ 表示所有数值变量的均值，$\overline{y_i}$ 表示名义变量 i 对应的所有数值变量的均值。*Eta* 系数的平方值越大，说明两变量的相关性越强。对 *V* 和 *Eta* 值来说，其值小于 0.3 时，相关性较弱；大于 0.6 时，相关性较强。虽然 *V* 和 *Eta* 值是不同指标，但指标的大小对应的相关程度是相同的。因此，当相关系数小于 0.3 时，均认为该特征与类别的相关性较小。特征与类别的相关性(Correlation coefficient, *CC*)公式表示为:

$$CC=\begin{cases}V & \text{f}\in\text{符号特征}\\ Eta & \text{f}\in\text{数值特征}\end{cases}. \tag{3}$$

相关性分析的结果将展示在实验部分。

入侵检测数据集不均衡性的特点使得入侵检测分类器不能有效的检测出数量较少的类型[33]，造成较高的误报率，因此，实现数据的均衡化对于提高异常检测效果是非常重要的步骤。不同的特征与类别相关性具有较大的差异，相关性强的特征通常来说具有更好的分类效果，因此，选择出信息量更多、相关性更强和分类效果更好的特征是非常有必要的。基于上述分析，本文对入侵检测数据集在均衡性和特征相关性两方面逐步进行优化。

3 基于流量异常分析多维优化的入侵检测方法

这一部分将详细描述本文提出的基于流量异常分析多维优化的入侵检测方法，对训练数据集和测试数据集进行预处理后，主要对数据的横向和纵向两个维度进行优化，详细过程如下所示:

(1) 数据横向优化过程: 提出了基于遗传算法的数据抽样优化算法。对训练数据集中数量较多的类别进行随机抽样，采用遗传算法优化各类别的抽样比例参数，准确性作为评价指标来评估抽样效果，得到一个均衡的训练数据集;

(2) 数据纵向优化过程: 提出了基于相关分析的特征选择优化算法。对训练数据集进行特征选择。结合特征与类别的相关分析，采用递归特征添加算法选择特征，提出平均召回率指标来评估特征选择效果，得到一个低维高效的训练数据集;

在(1)和(2)优化后的训练数据集上，采用随机森林模型构建入侵检测分类器。根据(2)提取测试集。在测试集上测试入侵检测分类器，得到异常检测结果。为了更好的展示本文方法，其整体流程如图 2 所示。

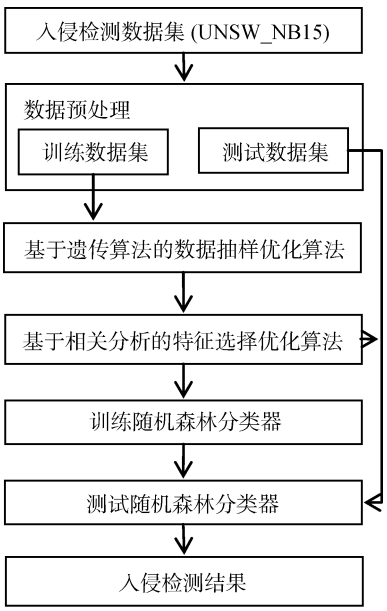


图 2 本文方法的整体流程图

Figure 2 The overall flow of the proposed method

随机森林算法用在(1)、(2)和最终的分类阶段，是因为该算法是一种集成方法，与其它算法相比，具有更好的分类效果。该算法以决策树为基础分类器，主要包括 3 个步骤:

(1) 选取训练集，构建基础分类器。采用自助采样的方法，给定包含 *m* 个样本的数据集，从数据集中有放回随机从中选取 m 次，得到跟原样本数量一致的数据集。若构造 *T* 个基础分类器，则需要 *T* 个新的数据集。

(2) 构建随机森林。假定当前一共有 *d* 个特征属性，从 *d* 个属性中随机选取 $k<d$ 个属性作为分类属性集。采用 CART(Classification and regression tree)方法构造单棵决策树，在 *k* 个属性中寻找能够在训练集中表现最优的属性进行分裂。参数 *k* 控制了特征的随机性，建议选择 $k=\log_2 d$。

(3) 投票。随机森林分类方法采用多数投票法进行决策。

随机森林算法能构建出更复杂的分类规则，对复杂数据有更好的分类效果[34]，并且该算法对参数的设置不敏感，可以很容易调整到一个合适的模型[34]。因此，随机森林分类器更适合数据量大、类别较多的异常检测数据集。

3.1 基于遗传算法的数据抽样优化算法

针对异常检测数据的不均衡问题，本文提出一种基于遗传算法的数据抽样优化算法。该算法的目标是对数量较多的类别进行抽样，通过减少数量较多的类别的数量，从而使整个数据集更为均衡。根据

上一部分提到的数据整体分布特点, 数据抽样分别在 Normal、Fuzzers、Reconnaissance、Exploit、DoS 和 Generic 这 6 类数量较多的类别上执行, 在每一类数据上设置一个抽样比例 r, 调整不同的抽样比例 r 可以得到不同数量的训练数据集, 最终会产生一组抽样比例组, 使得抽样的数据达到最优的分类效果。遗传算法被采用作为一种搜索策略, 综合的寻找最优的抽样比例组, 使新生成的训练数据集训练出更准确有效的分类器。

在该方法中, 采用随机抽样方法进行数据抽样。随机抽样法就是在总体中每个部分都有同等被抽中的可能, 是一种完全依照机会均等的原则进行的抽样行为。采用随机抽样的原因是该方法的简单性和高效性, 即较少的时间消耗和计算消耗。并且由于采用遗传算法对数据抽样比例参数进行搜索, 产生大量的解决方案, 这在一定程度上降低了随机抽样的随机性, 因此, 随机抽样方法是可行的。

遗传算法是一种随机搜索算法, 其主要操作包括适应性值计算和遗传操作两方面。适应性值表明个体或解的优劣性, 根据适应性值的大小进行接下来的遗传操作。遗传操作包括选择操作、交叉操作和变异操作。选择操作的目的是为了从当前群体中选出优良的个体, 使它们有机会作为父代为下一代繁殖子代。为体现这一思想, 选择操作使得适应性值高的染色体有更大的机会进入到下一代种群中, 实现了达尔文的适者生存原则。交叉操作是遗传算法中最主要的遗传操作。通过交叉操作产生新一代个体, 新个体保留了其父辈的特性。变异操作以一定的概率随机地改变某个染色体位置的值, 显然地, 变异发生的概率很低, 变异为新个体的产生提供了更多的机会。算法结束的终止条件是达到预先定义的迭代次数。在本文的参数优化过程中, 适应性值的计算和遗传操作的详细过程如算法 1 所示。

算法 1.基于遗传算法的数据抽样优化算法.

输入: 训练数据集 *Train data set*, 测试数据集 *Test data set*, 种群大小 N, 迭代次数 G

输出: 最优参数集合

$Up = \{1, 1, 1, 1, 1, 1\}$

$Low = \{0, 0, 0, 0, 0, 0\}$

$P = Random\_Generate\_Initial\_Population(N)$

FOR $i = 1{:}G$ do

Calculate *Fitness* of P

$Save\_Best\_Chromosome(P)$

P_{new} = empty

FOR $j = 1{:}\ N$ do

$Selection(P, X_{p1}, X_{p2})$

$X_0 = Crossover(X_{p1}, X_{p2})$

$Mutation(P_{new}, X_0)$

END FOR

$P = P_{new}$

END FOR

在该方法中染色体编码方式采用浮点型编码。虽然二进制编码方式是传统的更常用的编码方式, 但在该方法中真实的数值编码是更实用的, 这避免了二进制与十进制的频繁转变。每条染色体 $X=\{r_1, r_2, r_3, r_4, r_5, r_6\}$ 代表一个抽样方案, 用六元组表示, 其中 $r_i \in [0,1]$ 是第 i 类的抽样比例参数, 0 和 1 分别代表抽样比例参数的下限值和上限值。初始种群是随机生成的 N 条染色体, 每条染色体的值是 0~1 的随机数。初始种群生成后, 则计算每条染色体的适应性值, 根据适应性值进行选择、交叉和变异操作, 直到达到最大迭代次数。

一个染色体的适应性值决定这这条染色体在下一代种群中被选择的机会大小。在本方法中适应性值的计算是随机森林分类器的分类准确性。准确性(Accuracy, *Acc*)公式表示为:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

其中, *TP* (True positive)是指将异常样本正确分类为异常样本的数量, *TN* (True negative)是指将正常样本正确分类为正常样本的数量, *FP* (False positive)是指将正常样本错误分类为异常样本的数量, *FN* (False negative)是指将异常样本错误分类为正常样本的数量。*TP*、*TN*、*FP* 和 *FN* 的计算根据混淆矩阵得到的, 混淆矩阵公式如公式(5)所示。

$$H = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1l} \\ h_{21} & h_{22} & \cdots & h_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ h_{l1} & h_{l2} & \cdots & h_{ll} \end{bmatrix}, \tag{5}$$

其中, l 表示检测的类别数, h_{ii} 表示第 i 个类别正确检测为第 i 个类别的个数, h_{ij} 表示第 i 个类别被错误检测为第 j 个类别的个数。1 表示正常类别, 2~l 表示异常类别。*TP*、*TN*、*FP* 和 *FN* 的计算公式如下所示。

$$TP = \sum_{j=2}^{l} h_{jj} \tag{6}$$

$$TN = h_{11} \tag{7}$$

$$FP = \sum_{j=2}^{l} h_{1j} \tag{8}$$

$$FN = \sum_{j=2}^{l} h_{j1} \tag{9}$$

适应性值的计算结果将用于接下来的选择操作中。选择操作采用轮盘赌方法进行选择, 适应性值更高的染色体, 被选择进行下一代种群的机会越大。选择操作将选择两条染色体进入到交叉操作中。因为染色体的编码方式采用的是数值编码, 交叉和变异操作采用数值计算的方式进行。在交叉操作中, 在选择的两条染色体上, 随机选择两个基因位置, 计算两条染色体在该位置的平均值, 将平均值与适应值较高的染色体结合作为交叉操作的结果。交叉后的染色体进入到变异操作中。在变异操作中, 随机选择一个基因位置, 该位置有 0.5 的机会增加 0.01 或者减少 0.01。假设选择出两条染色体 X_1 和 X_2, X_1 有更高的适应性值, X_1 和 X_2 交叉操作和变异操作如图 3 所示。

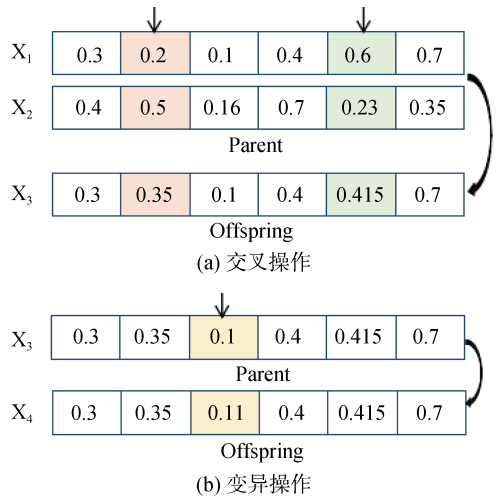


图 3　遗传操作中的交叉和变异操作

Figure 3　Crossing and mutation operation in genetic operation

3.2　基于相关分析的特征选择优化算法

这一部分将详细介绍本文提出的基于相关分析的特征选择优化算法。该算法包含两个步骤: 特征与类别的相关分析和递归特征添加算法。特征与类别的相关分析采用第二章介绍的相关分析方法, 计算特征与类别的相关系数后, 删除相关系数小于 0.3 的特征, 得到一个与类别相关性强的特征子集。该特征子集应用于接下来进行递归特征添加算法, 该算法每次添加一个特征, 使已选择的特征子集达到最优的分类效果。

递归特征添加算法是一种包装式特征选择算法, 与某种分类模型和特征选择效果评价方法相结合, 最终选取评价效果最好的特征子集。递归特征添加是一种特征选择的贪婪搜索策略, 在该方法中, 初始化一个空的特征集合, 逐步向其中添加新的特征, 如果该特征能提高预测效果, 即得以保留, 否则就扔掉。值得注意的是, 该方法针对每一个特征子集重新训练模型, 具有较大的计算量。因此, 针对这一问题, 本文将特征的相关分析与递归特征添加算法相结合, 通过一个相关性较强的特征子集来减少递归特征添加算法的计算量。该方法保证了特征子集的分类效果和选择效率。

本文方法采用随机森林作为分类器, 并提出平均召回率作为评价方法。首先初始化一个空集, 用于存放将要选择的特征。接下来从剩余的特征中每次递归的添加一个特征, 与已选择的特征结合而产生的最优平均召回率的特征会添加到已选择的特征集合中, 最终所有的特征均被选择。随着特征子集的增大, 分析已选特征子集的分类效果, 当特征子集的分类效果不在增加时, 表明此时已选的特征具有最优的平均召回率, 该特征子集为最终选择的特征子集。详细的算法过程如算法 2 所示。

算法 2.基于相关分析的特征选择优化算法.

输入: 训练数据集 *Train data set*, 测试数据集 *Test Data set*, 特征集 *F*, 特征集的个数 *N*

输出: 选择的特征子集 *SF*

Calculate *CC* of features with label

Get a new features sub-set *new_F* by deleting features of $CC \leqslant 0.3$

Set an empty selected features set *F'*

Set an empty max *Avg_Recall* set *R*

FOR $i = 1{:}N$ do

　　FOR each $f \in new\_F$ do

　　Calculate *Avg_Recall* of $\{F' \cup f\}$ with Random Forests Algorithm

　　End FOR

　　Get *f* of *max Avg_Recall*

　　F'.add(*f*)

　　R.add(*max Avg_Recall*)

　　new_F.remove(*f*)

END FOR

WHILE $R_i > R_{i-1}$ do

　　SF. add(F'_i)

END WHILE

算法 2 中的平均召回率指标 *Avg_Recall*(Average Recall), 即各类别召回率的平均值。平均召回率指标的提出是为了有效的评估各个类别的检测率, 尤其是对数量较少的类别的检测。该指标是基于公式(5)得出的, 第 *i* 个类别的召回率如下所示。

$$\mathrm{Re}call_i = \frac{h_{ii}}{\sum_{j=1}^{l} h_{ij}}, \tag{10}$$

平均召回率计算方式如(11)所示。

$$Avg\_\mathrm{Re}call = \frac{\sum_{i=1}^{l}\mathrm{Re}call_i}{l} \tag{11}$$

该方法的最终结果是获取一个平均召回率最高的特征子集。在该方法中单独特征的分类效果不在是关注的重点, 关注的重点是特征集合或者是关联特征的分类重要性。该方法的主要目标在于消除特征冗余, 保证特征与类别的相关性和特征子集最优的分类效果。

为了更好的说明递归特征添加算法, 本文给出一个实例来说明递归特征添加的详细过程, 如图 4 所示。假设该数据有 3 个特征$\{f_1, f_2, f_3\}$, 每个特征有两种状态 0 和 1, 0 表示该特征没有被选择, 1 表示该特征已被选择。初始状态时, 3 个特征均没有被选择, 状态表示为{0, 0, 0}; 接下来逐次计算每个特征的平均召回率, 特征f_2的平均召回率最高, 则特征f_2被选择; 接下来在特征f_2的基础上, 分别计算特征组合$\{f_1, f_2\}$和$\{f_2, f_3\}$的平均召回率, 特征组合$\{f_1, f_2\}$的平均召回率最高, 则特征f_1被选择。在最终状态时, 全部特征被选择。图 3 中的实线表示特征选择的过程, 该实例中特征选择的顺序为$\{f_2, f_1, f_3\}$。值得注意的是, 随着选择特征的增多, 特征子集的平均召回率并不是持续提高的, 也就是说特征子集$\{f_2\}$的平均召回率并不一定低于特征子集$\{f_1, f_2\}$的平均召回率。

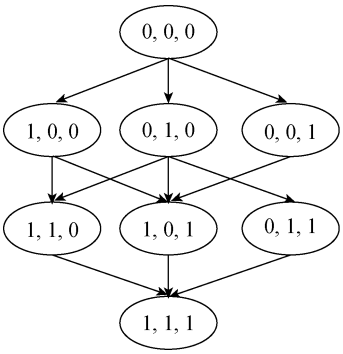


图 4 递归特征添加算法实例

Figure 4 The example of recursive feature addition algorithm

4 实验结果

在这一部分, 我们将会评估本文提出算法的性能, 所有的实验均是在 Windows 7 PC, Intel(R) Xeon(R) CPU E5-2603 0 @1.80GHz 1.80GHz and 8.00GB RAM 环境中实现。采用 Python 3.5.2 实现本文的算法。

4.1 训练集和测试集

UNSW-NB15 数据集用于评价本文提出的算法。UNSW-NB15 数据集[35]是由澳大利亚安全实验室采用 IXIA PerfectStrom tool 生成的, 该数据结合现代网络中真实的正常网络流量和人造的攻击流量。该数据集的特征分为 4 类: 基本特征、内容特征、时间特征和额外生成的特征, 详细的特征描述如表 1 所示。数据集中攻击类型包括 9 类: Fuzzers、Analysis、Backdoor、DoS、Exploit、Generic、Reconnaissance、Shellcode、Worm, 详细的攻击行为描述和数量如表 2 所示。

4.2 数据集预处理

训练数据集和测试数据集在应用到异常检测方法之前, 需要对数据进行预处理。数据预处理过程共包括 2 个步骤: (1) 将训练集和测试集中的符号特征 protocol, service, state 转化为数值表示。以特征 state 为例, 该属性共包括 11 个变量, 用数值 1~11 依次表示 11 个变量。(2) 将类别标签转化为数值表示, 其中 1 表示 Normal 类别, 2 表示 Backdoor 类别, 3 表示 Analysis 类别, 依次类推。

4.3 实验评估指标

除了在第三章部分提到的准确性 *Acc* 和召回率 *Recall* 之外, 本文还采用检测率 *DR* (Detection rate) 和误报率 *FAR* (False alarm rate)作为入侵检测方法的评估指标[36]。

$$DR = \frac{TP}{TP + FN}, \tag{12}$$

表 1 UNSW-NB15 数据集的特征描述

Table 1 Features description of UNSW-NB15 data set

特征类别	特征名称
基本特征	dur(1), proto(2), service(3),state(4), spkts(5), dpkts(6), sbytes(7), dbytes(8), rate(9), sttl(10), dttl(11), sload(12), dload(13), sloss(14), dloss(15)
内容特征	swin(16), dwin(17), stcpb(18), dtcpb(19), smean(20), dmean(21), trans_depth(22), res_bdy_len(23)
时间特征	sintpkt(24), dintpkt(25), sjit(26), djit(27), tcprtt(28), synack(29), ackdat(30)
额外生成的特征	ct_srv_src(31), ct_state_ttl(32), ct_dst_ltm(33), ct_src_dport_ltm(34), ct_dst_sport_ltm(35), ct_dst_src_ltm(36), is_ftp_login(37), ct_ftp_cmd(38), ct_flw_http_mthd(39), ct_src_ltm(40),ct_srv_dst(41),is_sm_ips_ports(42)

(注: 特征名称的数字标号并没有实际意义, 后文中将使用数字标号代表特征。)

表 2 UNSW_NB15 的攻击行为和攻击数量描述

Table 2 The description of Attack behavior and number of UNSW-NB15 data set

名称	行为描述	训练集数量	测试集数量
Normal		56000	37000
Backdoor	绕过安全性控制而获取对程序或系统访问权的技术。	1746	583
Analysis	一种通过端口、电子邮件和 web 脚本渗透 web 应用程序的入侵方式。	2000	677
Fuzzers	一种试图发现程序、操作系统或网络中安全漏洞的攻击方式, 通过输入大量随机数据, 使其崩溃。	18184	6062
Shellcode	一种通过发送利用特定漏洞的代码控制目标机器的攻击方式。	1133	378
Reconnaissance	一种为逃避安全控制而收集计算机网络信息的攻击方式。	10491	3496
Exploit	一段通过触发一个漏洞(或者几个漏洞)进而控制目标系统的代码。	33393	11132
DoS	一种通过直接或间接耗尽被攻击对象的资源, 使目标计算机或网络无法提供正常的服务或资源访问的攻击方式。	12264	4089
Worm	一种通过网络传播的主动攻击的恶性计算机病毒。	130	44
Generic	一种不考虑分组密码的配置而使用哈希函数对每个分组密码进行碰撞的技术。	40000	18871
Total		175341	82332

$$FAR = \frac{FP}{TN + FP}. \tag{13}$$

4.4 实验分析

第一个实验用来评估本文提出的数据抽样优化算法的有效性。通过比较数据抽样前后, 算法的分类准确性、检测率和误报率来说明本文提出的数据抽样方法的有效性。在基于遗传算法的数据抽样优化算法中, 通过不断地调整算法中的种群大小 N 和迭代次数 G, 在 N=10, G=30 时, 实现的最优的准确率为 0.911, 此时的抽样比例参数为{0.876, 0.048, 0.639, 0.448, 0.706, 0.920}。表 3 直观的展示了数据抽样前后各个类别的数据量, 与原始数据相比, 数据抽样后各个类别的数据分布更为均衡。图 5 展示了原始数据与抽样后的数据分别作为训练集来训练随机森林分类器, 在整个测试集进行测试, 得到的准确率、检测率和误报率。采用原始数据训练分类器时, 具有较好的检测率, 但误报率偏高。数据抽样后, 检测准确率明显提高, 误报率显著降低, 检测率略微降低。

数据抽样前后, 每个类别的召回率结果如表 4 所示。异常行为 Backdoor 和 Analysis 在数据抽样前后的召回率都普遍较低, 这是因为这两类攻击行为本身数量非常少, 且与正常行为较为类似, 从而导致这两种行为难以检测。数据抽样后, Fuzzers 类型的召回率降低, Exploits 类型的召回率略微降低, 这是因为 Fuzzers 攻击行为在数据抽样后, 数据量大量减少, 导致其召回率出现减少。而其他类型的召回率均有明显提高。

表 3 数据抽样前后各个类别的数据量

Table 3 The number of each category before and after data sampling

类别	数量 / 数据抽样前	数量 / 数据抽样后
Normal	56000	49068
Backdoor	1746	1746
Analysis	2000	2000
Fuzzers	18184	868
Shellcode	1133	1133
Reconnais-sance	10491	6699
Exploits	33393	14944
DoS	12264	8661
Worm	130	130
Generic	40000	6785

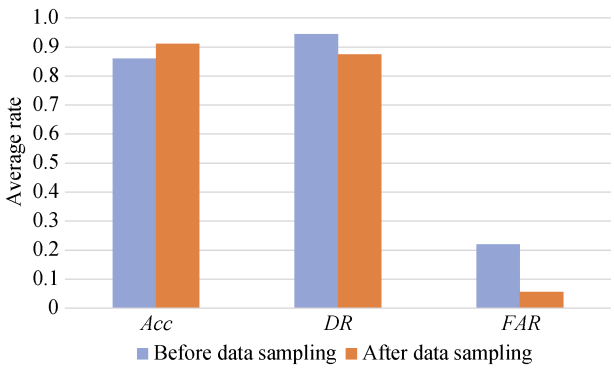


图 5 数据抽样前后异常检测效果对比

Figure 5 Comparison of anomaly detection effect before and after data sampling

第二个实验用来评估本文提出的基于相关分析的特征选择优化算法的有效性。通过比较全部特征和最终选择的特征子集的分类效果来说明本文提出的

特征选择优化算法能提高异常行为的检测效果。特征与类别相关性的计算结果如表 5 所示。结果显示, 共有 7 个特征的相关性小于 0.3, 将这 7 个特征数据删除, 剩余的 35 个特征用于接下来的递归特征添加算法。

表 4　数据抽样前后各个类别的召回率对比

Table 4　Comparison of Recall of each category before and after data sampling

类别	数据抽样前	数据抽样后
Normal	0.780	**0.941**
Backdoor	0.069	**0.072**
Analysis	**0.068**	0.055
Fuzzers	**0.542**	0.069
Shellcode	0.648	**0.783**
Reconnais-sance	0.800	**0.811**
Exploits	**0.723**	0.647
DoS	0.129	**0.234**
Worm	0.182	**0.227**
Generic	0.968	**0.970**

表 5　特征与类别的相关性

Table 5　The correlation coefficient (*CC*) between features with label

特征	*CC*	特征	*CC*
1	0.873	40	0.599
20	0.834	30	0.594
7	0.815	24	0.584
12	0.790	15	0.545
9	0.748	14	0.530
32	0.745	26	0.524
11	0.735	27	0.515
35	0.719	16	0.471
21	0.716	18	0.461
6	0.703	19	0.461
10	0.701	17	0.459
13	0.686	3	0.323
36	0.666	2	0.319
41	0.661	**4**	**0.260**
34	0.655	**42**	**0.157**
31	0.644	**23**	**0.101**
33	0.637	**39**	**0.049**
8	0.635	**22**	**0.041**
25	0.631	**37**	**0.026**
5	0.613	**38**	**0.026**
28	0.599		
29	0.599		

基于相关分析的特征选择优化算法的最终结果是得到一个平均召回率最高的特征子集。所选特征子集的数量与平均召回率的关系如图 6 所示。在特征选择初期, 随着选择的特征的增多, 平均召回率逐渐升高; 在选择 11~13 个特征时, 平均召回率达到最高, 约为 0.63; 接下来, 随着选择特征的增多, 平均召回率逐步下降。在选择一个特征与选择全部特征时, 平均召回率的结果基本相同。此结果有效地说明了特征选择的重要作用, 当特征较少时, 入侵检测分类器不能有效地对异常行为进行分类, 当全部特征用于训练入侵检测分类器时, 入侵检测分类器的作用同样会受到较大的影响。在该方法中, 最终选择的特征的个数为 12, 选择的特征子集为{7, 8, 10, 36, 41, 21, 3, 15, 5 , 6, 20, 13}, 此时的平均召回率最高。

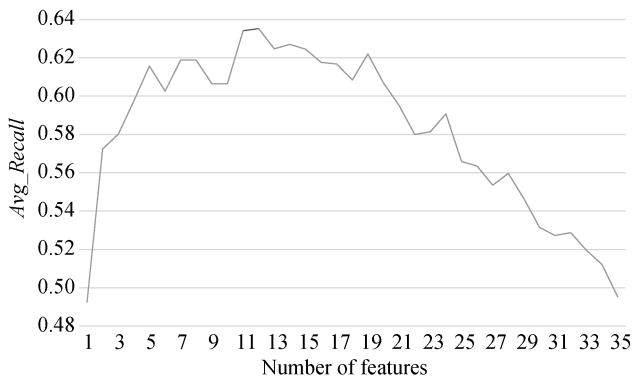


图 6　特征选择的数量与平均召回率之间的关系

Figure 6　The relationship between the number of selected features and the average Recall

将全部特征的检测效果与所选择特征子集的准确率、检测率和误报率进行对比, 结果如表 6 所示。在特征选择前后, 检测效果的准确率、检测率和误报率基本上保持不变, 这与特征选择的理论作用是相吻合的。在本文中, 特征选择的作用主要体现在对各个类别的检测率上。图 7 展示了各个类别的召回率结果, 结果显示各个类别的检测效果有显著的变化。除了 Normal 类别的召回率略微降低外, 其他类别的召回率均有显著提高或保持不变。因此, 基于相关分析的特征选择优化算法是有效的。

通过上述分析, 说明本文提出的方法在数据的横向维度和纵向维度逐步优化, 入侵检测的效果逐步提升, 总体检测效果和各个类别的检测效果如图 8 和图 9 所示, 异常检测结果的准确性逐步提高, 误报率取得了显著降低。在各个类别的召回率上, 除了 Fuzzers 类别的召回率不太理想, Exploits 类别的召回率略微降低之外, 其他类别的召回率均有所提高,

表 6 特征选择前后检测效果对比

Table 6 Comparison of anomaly detection effect before and after features selection

评估指标	特征选择前	特征选择后
Acc	0.911	0.906
DR	0.875	0.883
FAR	0.056	0.072

尤其是 Analysis、DoS 和 Worm 类别的召回率出现显著提高。总而言之，本文的优化过程是行之有效的。

为了更好地验证本文提出的算法，将本文提出的算法与其他算法在准确率、误报率以及各个类别的召回率上进行对比，表 7 和图 10 分别展示了比较结果。表 7 展示了本文算法与其他算法在准确率和误报率的对比结果。本文提出的算法的准确率为 90.6%，相比于其他算法，高出约 10%左右，检测效果最优，且误报率远远低于 Moustafa 等人提出的算法，仅仅比 GALR_DT 算法的误报率高出 1%左右。除此之外，本文方法最终选择的特征为 12 个，远远少于其他算法选取的特征数。因此，本文算法在检测效率上是最高的，其检测时间和计算消耗要远远少于其他算法。

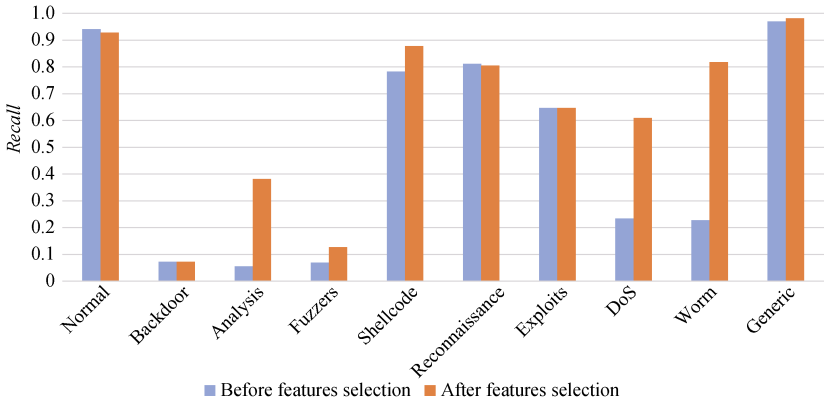


图 7 特征选择前后各个类别的召回率对比

Figure 7 Comparison of Recall of each category before and after features selection

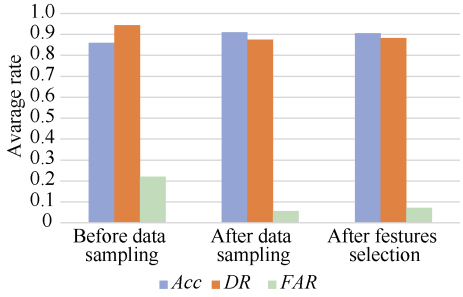


图 8 本文算法的入侵检测效果

Figure 8 Intrusion detection result of the proposed algorithm

图 10 展示了本文算法与 GALR_DT 算法在各个类别的召回率上的比较结果，除了 Fuzzers 和 Exploits 这两个类别外，其他类别的召回率均有显著提高，尤其是类别 Analysis、Shellcode、DOS 和 Worm 的召回率。通过上述分析，Analysis、Shellcode、DOS 和 Worm 这 4 中攻击类型均属于数据量较少的攻击行为，本文的优化过程对比结果(图 9)和与其他算法的对比结果(图 10)来看，本文方法对数量较少的攻击类别的召回率有显著的影响。综上所述，本文算法在异常检测中是有效可行的。

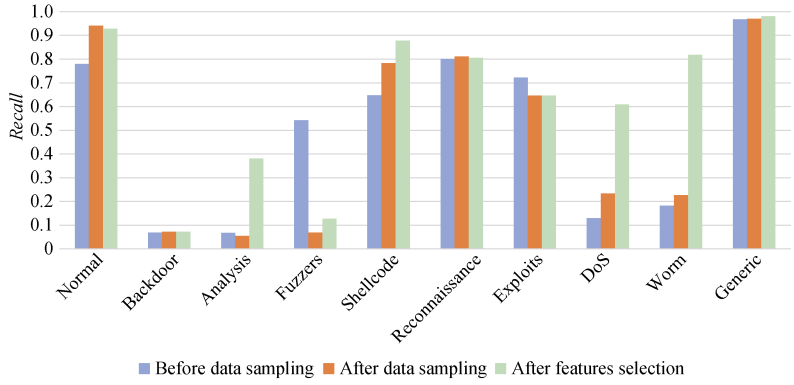


图 9 本文算法的各个类别的召回率结果

Figure 9 Recall result of the proposed method

表 7　本文提出算法与其他算法的检测效果对比

Table 7　Comparison of anomaly detection effect between the proposed algorithm and other algorithms

方法	特征数量	分类器	*Acc*	*FAR*
Moustafa and Slay[32]	42	DT	0.8556	0.1578
		LR	0.8315	0.1848
		NB	0.8207	0.1856
		ANN	0.8134	0.2113
		EM	0.7847	0.2379
GALR_DT[17]	20	DT	0.8142	0.0639
提出的算法	12	RF	**0.9060**	**0.0720**

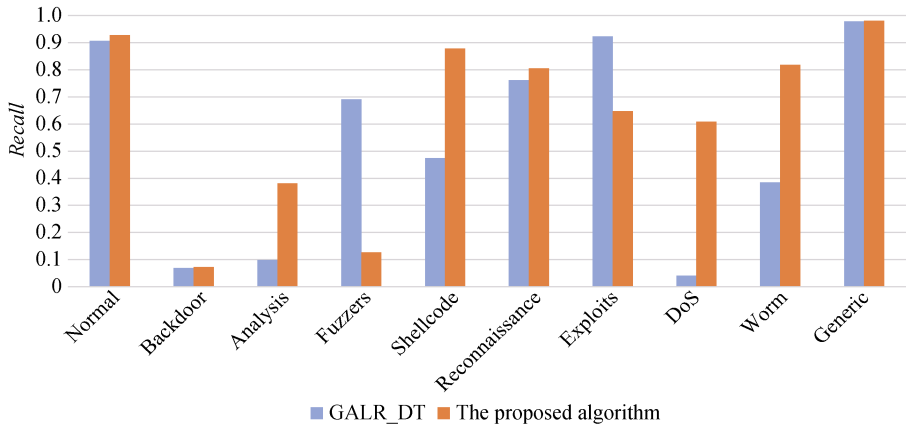


图 10 本文算法与 GALR_DT 的召回率对比结果

Figure 10　Comparison of Recall between the proposed algorithm and GALR_D

5　结论

为了更有效的检测出网络异常行为, 本文提出了基于流量异常分析多维优化的入侵检测方法, 以统计分析为基础, 从数据的横向和纵向两个维度进行优化, 在横向维度上提出了基于遗传算法的数据抽样优化算法, 在纵向维度上提出了基于相关分析的特征选择优化算法。在基于遗传算法的数据抽样优化算法中, 采用随机抽样算法对多数类进行数据抽样, 遗传算法作为搜索策略, 用来优化每个类别中的数据抽样比例参数, 得到最优的比例参数来抽样数据集, 得到一个数量少、类别均衡的训练数据集, 该训练数据集提高了入侵检测的准确性, 并降低了误报率。在此基础上, 执行基于相关分析的特征选择优化算法。该算法首先删除相关性小的特征, 得到一个相关性强的特征子集; 基于该特征子集, 执行递归特征添加算法, 以平均召回率作为评价指标, 最终得到一个平均召回率最高的特征子集。该特征子集能有效地提高每个类别的召回率, 得到更有效的异常检测结果。最后采用随机森林模型作为入侵检测的分类器, 在均衡低维的训练数据集进行训练, 整个测试集进行测试, 得到最终的异常检测结果。真实数据集 UNSW_NB15 用来评估本文提出的方法。通过逐步执行本文方法, 异常检测的准确率逐步提高, 误报率明显降低, 并且不同类别的召回率有明显的改善。与其他算法相比, 本文提出的算法有一个高准确率和低误报率, 并在每个类别取得了有效的召回率。

以后的研究工作将考虑结合欠采样和过采样方法, 进一步解决入侵检测数据的不均衡问题, 从而更有效的提高入侵检测的准确性, 降低误报率。并并考虑在多个入侵检测数据集上进行实验, 从而更全面的验证所提出算法的效果。

致　谢　感谢国家重点研发计划项目 (No. 2016YFB0800700), 国家自然科学基金 (No. 61472341, No. 61772449, No. 61572420, No. 61807028, No. 61802332), 河北省自然科学基金(No. F2016203330), 博士后科研择优资助项目 (No. B2017003005)的资助。特别感谢任家东教授、何海涛教授和王倩讲师的指导, 感谢郭嘉伟同学在实验上的帮助, 感谢实验室同学的帮助。

参考文献

[1] H. Pajouh, G. Dastghaibyfard, and S. Hashemi, “Two-tier network anomaly detection model: a machine learning approach,” *Journal of Intelligent Information Systems*, vol. 48, no. 1, pp. 1-14, 2015.

[2] A. Amaral, L. Mendes, B. Zarpelao, and M. Junior, “Deep IP flow inspection to detect beyond network anomalies,” *Computer Communications*, vol. 98, pp. 80-96, Jan. 2017.

[3] CF. Tsai, YF. Hsu, CY. Lin, and WY. Lin, “Intrusion detection by machine learning: A review,” *Expert Systems with .Application*, vol. 36, no. 10, pp.11994-12000, 2009.

[4] YY. Chung, and N. Wahid, “A hybrid network intrusion detection system using simplified swarm optimization(SSO),” *Applied Soft Computing Journal*, vol. 12, no. 9, pp. 3014-3022, 2012.

[5] B. Jain, “Intrusion Prevention and Vulnerability Assessment in BCEFHP Intrusion Detection System [Master.dissertation],” Indian Institute of Technology, Kanpur, 2005.

[6] D. Kwon, H. Kim, J. Kim, S. Suh, and K. Jim, “A survey of deep learning-based network anomaly detection,” *Cluster Computing*, vol. 20, no. 3, pp. 1-13, Sep. 2017.

[7] W.H. Chen, S.H. Hsu, and H.P. Shen, “Application of SVM andANN for intrusion detection,” *Computers & Operations Research*, vol. 32, no. 10, pp. 2617-2634, 2005.

[8] P. Sangmee, N. Thanon, and N. Elz, “Anomaly detection using new MIB traffic parameters based on profile,” in Proc. *Computing Technology and Information Management* (ICCM’ 8), pp. 648–653, 2012.

[9] Q. Kang and Q. D. Wu, “Imbalance Classification Methods in Machine Learning,” Tongji University Press, 2017.
(康琦, 吴启迪, “机器学习中的不平衡分类方法”, 同济大学出版社, 2017).

[10] W. Al-Yaseen, Z. Othman, and M. Nazri, “Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system,” *Expert Systems with Applications*, vol. 67, pp. 296-303, 2017.

[11] W. Al-Yaseen, Z. Othman, and M. Nazri, “Intrusion Detection System Based on Modified K-means and Multi-level Support Vector Machines,” in Proc. *International Conference on Soft Computing in Data Science*, pp. 265-274, 2015.

[12] W. Al-Yaseen, Z. Othman, and M. Nazri, “Hybrid Modified K-Means with C4.5 for Intrusion Detection Systems in Multiagent Systems,” *Scientific world journal*, vol. 2015, no. 2, 2015.

[13] AS. Eesa, Z. Orman, and AMA. Brifcani, “A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems,” *Expert Systems with Applications*, vol. 42, no. 5, pp. 2670-2679, Nov. 2015.

[14] M. Bennasar, Y. Hicks, and R. Setchi, “Feature selection using Joint Mutual Information Maximisation,” *Expert Systems with Applications*, vol. 42, no. 22, pp. 8520-8532, Dec. 2015.

[15] W. Gao, L. Hu, P. Zhang, and F. Wang, “Feature selection by integrating two groups of feature evaluation criteria,” *Expert Systems with Applications*, vol. 110, pp. 11-19, Nov. 2018.

[16] R. Alshboul, F. Thabtah, N. Abdelhamid, and M. Al-Diabat, “A Visualization Cybersecurity Method based on Features’ Dissimilarity,” *Computers & Security*, vol. 77, pp. 289-303, Aug. 2018.

[17] C. Khammassi, and S. Krichen, “A GA-LR Wrapper Approach for Feature Selection in Network Intrusion Detection,” *Computers & Security*, vol. 70, pp. 255-277, 2018.

[18] EDL. Hoz, EDL. Hoz, A. Ortiz, J. Ortega, and AM. Alvarez, “Feature selection by multi-objective optimisation: Application to network anomaly detection by hierarchical self-organising maps,” *Knowledge-Based Systems*, vol. 71, pp. 322-338, Nov. 2014.

[19] G. Herman, B. Zhang, Y. Wang, G. Ye, and F. Chen, “Mutual information-based method for selecting informative feature sets,” *Pattern Recognition*, vol. 46, no. 12, pp. 3315-3327, Dec. 2013.

[20] M. Bennasar, Y. Hicks, and R. Setchi, “Feature selection using Joint Mutual Information Maximisation,” Pergamon Press, 2015.

[21] S. Aljawarneh, M. Aldwairi, and MB. Yassein, “Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model,” *Journal of Computational Science*, vol. 25, pp. 152-160, Mar. 2017.

[22] T. Hamed, R. Dara, and SC. Kremer, “Network intrusion detection system based on recursive feature addition and bigram technique,” *Computers & Security*, vol. 73, Nov. 2017.

[23] M. Ahmed, AN. Mahmood, and J. Hu, “A survey of network anomaly detection techniques,” *Journal of Network & Computer Applications*, vol. 60, pp. 19-31, Jan. 2016.

[24] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, “Anomaly-based network intrusion detection: Techniques, systems and challenges,” *Computers & Security*, vol. 28, no. 1, pp. 18-28, Feb. 2009.

[25] AA. Ghorbani, W. Lu, and M. Tavallaee, “Network intrusion detection and prevention: concepts and techniques,” *Advances in Information Security*, vol. 28, no. 3, pp. 42-48, 2012.

[26] C. Modi, D. Patel, B. Borisaniya B, and et al, “Review: A survey of intrusion detection techniques in Cloud,” *Journal of Network & Computer Applications*, vol. 36, no. 1, pp. 42-57, Jan. 2013.

[27] W. N. Lin, M. Z. Chen, Y. Q. Zhan, and C.B. Liu, “Research on an intrusion detection algorithm based on PCA and Random-forest classification,” *Netinfo Security*, vol. 11, pp. 50-54, 2017.
(林伟宁, 陈明志, 詹云清, 刘川葆. “一种基于PCA和随机森林分类的入侵检测算法研究”, *信息网络安全*, 2017, 11: 50-54.)

[28] CF. Tsai, YF. Hsu, CY. Lin, and WY. Lin, “Intrusion detection by machine learning: a review,” *Expert Systems with Applications*, vol. 36, no. 10, pp. 11994-12000, Dec. 2009.

[29] AA. Aburomman , and M B I.Reaz, “A Novel Weighted Support Vector Machines Multiclass Classifier Based on Differential Evolution for Intrusion Detection Systems,”. *Information Sciences*, vol. 414, pp. 225-246, 2017.

[30] U. Fiore, F. Palmieri, A. Castiglione, and AD. Santis, “Network anomaly detection with the restricted Boltzmann machine,” *Neurocomputing*, vol. 122, pp. 13-23, Dec. 2013.

[31] R. Vijayanand, D. Devaraj, and B. Kannapiran, “Intrusion detection system for wireless mesh network using multiple support vector machine classifiers with genetic-algorithm-based feature selection,” *Computers & Security*, vol. 77, pp. 304-314, Aug. 2018.

[32] N. Moustafa N, and J. Slay, “The evaluation of Network Anomaly

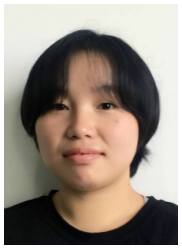
Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set," *Information Systems Security*, vol. 25, no. 1, pp. 18-31, Jan. 2016.

[33] YT. Guo, Y. Gao, Y. Wang, and et al, "DPI & DFI: A Malicious Behavior Detection Method Combining Deep Packet Inspection and Deep Flow Inspection," *Procedia Engineering*, vol. 174, pp. 1309-1314, 2017.

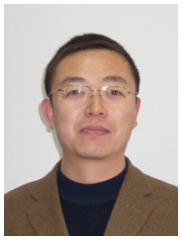
[34] Leo B, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[35] https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets, 2015

[36] [MRG. Raman, N. Somu, K. Kirthivasan, and et al, "An efficient intrusion detection system based on hypergraph-genetic algorithm for parameter optimization and feature selection in support vector machine," *Knowledge-Based Systems*, vol. 134, pp. 1-12, Oct. 2017.



刘新倩 于 2015 年在燕山大学教育技术学专业获得学士学位。现在燕山大学计算机科学与技术专业攻读博士学位。研究领域为网络安全、数据挖掘。研究兴趣包括: 网络安全度量、异常检测。Email: 1689890718@qq.com



任家东 于 1999 年在哈尔滨工业大学计算机应用技术专业获得博士学位。现任燕山大学信息科学与工程学院院长, 教授, 博士生导师。研究领域为数据挖掘、数据建模、软件安全。研究兴趣包括: 软件安全、网络安全。Email: jdren@ysu.edu.cn

郭嘉伟 于 2016 年在石家庄学院计算机科学与技术专业获得学士学位。现于燕山大学计算机专业攻读硕士学位。研究领域为网络安全、数据挖掘。研究兴趣包括: 网络安全度量、异常检测。Email：739799132@qq.com

单纯 于 2015 年在北京理工大学计算机应用技术专业获得博士学位。现任北京理工大学软件理论研究所讲师。研究领域为软件安全、人工智能、软件测试与质量保证。研究兴趣包括: 人工智能、软件测试与质量保证。Email: sherryshan@bit.edu.cn

王倩 于 2016 年在燕山大学计算机软件与理论专业获得博士学位。现任燕山大学信息科学与工程学院讲师。研究领域为数据挖掘、软件安全、网络安全。研究兴趣包括: 软件安全、网络安全。Email: wangqianysu@163.com