计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20180063
56(3):566-575,2019

# 基于 *K*NN 离群点检测和随机森林的多层入侵检测方法

任家东[1,2]　刘新倩[1,2]　王　倩[1,2]　何海涛[1,2]　赵小林[3,4]
[1](燕山大学信息科学与工程学院　河北秦皇岛　066001)
[2](河北省软件工程重点实验室(燕山大学)　河北秦皇岛　066001)
[3](北京理工大学软件学院　北京　100081)
[4](软件安全工程技术北京市重点实验室(北京理工大学)　北京　100081)
(jdren@ysu.edu.cn)

# An Multi-Level Intrusion Detection Method Based on *K*NN Outlier Detection and Random Forests

Ren Jiadong[1,2], Liu Xinqian[1,2], Wang Qian[1,2], He Haitao[1,2], and Zhao Xiaolin[3,4]
[1](*School of Information Science and Engineering, Yanshan University, Qinhuangdao, Hebei* 066001)
[2](*Hebei Key Laboratory of Software Engineering (Yanshan University), Qinhuangdao, Hebei* 066001)
[3](*School of Software, Beijing Institute of Technology, Beijing* 100081)
[4](*Beijing Key Laboratory of Software Security Engineering Technology (Beijing Institute of Technology), Beijing* 100081)

**Abstract**　Intrusion detection system can efficiently detect attack behaviors, which will do great damage for network security. Currently many intrusion detection systems have low detection rates in these abnormal behaviors Probe (probing), U2R (user to root) and R2L (remote to local). Focusing on this weakness, a new hybrid multi-level intrusion detection method is proposed to identify network data as normal or abnormal behaviors. This method contains *K*NN (*K* nearest neighbors) outlier detection algorithm and multi-level random forests (RF) model, called *K*NN-RF. Firstly *K*NN outlier detection algorithm is applied to detect and delete outliers in each category and get a small high-quality training dataset. Then according to the similarity of network traffic, a new method of the division of data categories is put forward and this division method can avoid the mutual interference of anomaly behaviors in the detection process, especially for the detecting of the attack behaviors of small traffic. Based on this division, a multi-level random forests model is constructed to detect network abnormal behaviors and improve the efficiency of detecting known and unknown attacks. The popular KDD (knowledge discovery and data mining) Cup 1999 dataset is used to evaluate the performance of the proposed method. Compared with other algorithms, the proposed method is significantly superior to other algorithms in accuracy and detection rate, and can detect Probe, U2R and R2L effectively.

**Key words**　network security; intrusion detection system; *K*NN outlier detection; random forests model; multi-level

**收稿日期**:2018-01-26;**修回日期**:2018-06-05

**基金项目**:国家重点研发计划基金项目(2016YFB0800700);国家自然科学基金项目(61472341,61772449,61572420);河北省自然科学基金项目(F2016203330, F2015203326);燕山大学博士后科研择优资助项目(B2017003005);燕山大学博士基金项目(B1036)
This work was supported by the National Key Research and Development Program of China (2016YFB0800700), the National Natural Science Foundation of China (61472341, 61772449, 61572420), the Natural Science Foundation of Hebei Province of China (F2016203330, F2015203326), the Advanced Program of Postdoctoral Scientific Research of Yanshan University (B2017003005), and the Doctoral Foundation of Yanshan University (B1036).

**通信作者**:王倩(wangqianysu@163.com)

**摘 要** 入侵检测系统能够有效地检测网络中异常的攻击行为,对网络安全至关重要.目前,许多入侵检测方法对攻击行为 Probe(probing),U2R(user to root),R2L(remote to local)的检测率比较低.基于这一问题,提出一种新的混合多层次入侵检测模型,检测正常和异常的网络行为.该模型首先应用 KNN(K nearest neighbors)离群点检测算法来检测并删除离群数据,从而得到一个小规模和高质量的训练数据集;接下来,结合网络流量的相似性,提出一种类别检测划分方法,该方法避免了异常行为在检测过程中的相互干扰,尤其是对小流量攻击行为的检测;结合这种划分方法,构建多层次的随机森林模型来检测网络异常行为,提高了网络攻击行为的检测效果.流行的数据集 KDD(knowledge discovery and data mining) Cup 1999 被用来评估所提出的模型.通过与其他算法进行对比,该方法的准确率和检测率要明显优于其他算法,并且能有效地检测 Probe,U2R,R2L 这 3 种攻击类型.

**关键词** 网络安全;入侵检测系统;KNN 离群点检测;随机森林模型;多层次

**中图法分类号** TP393.08

随着计算机和网络服务的不断发展,计算机和网络的安全已经成为一个重要的问题.网络中异常行为的检测已经成为网络安全领域重要的研究内容.入侵检测系统用来检测和分析网络数据,标识异常的网络行为.通常来讲,入侵检测系统主要分为 2 类:基于误用的入侵检测系统和基于异常的入侵检测系统[1].基于误用的入侵检测系统可以有效地检测出已知的攻击类型,例如 Snort 入侵检测系统[2].这种类型的入侵检测系统误报率较低,但是不能很好地识别网络中新的攻击类型.基于异常的入侵检测系统能够根据正常的网络行为建立检测模型,识别出正常行为的偏差值,从而识别网络异常行为[3].这种类型的入侵检测系统能够检测出新的或未知的攻击类型,但又有较高的误报率.

为了降低基于异常的入侵检测系统的误报率,许多数据挖掘和机器学习方法,例如支持向量机(support vector machine, SVM)和神经网络,被应用到入侵检测系统中.目前的许多研究将特征选择或特征提取的方法与一些机器学习的方法相结合,来提高入侵检测系统的准确性,同时降低算法的运行时间.Raman 等人[4]提出将超图、遗传算法和支持向量机相结合来实现入侵检测系统.超图和遗传算法用于实现支持向量机模型的参数估计和特征选择,支持向量机用来对特征选择后网络数据进行异常检测,证明了特征选择方法和支持向量机结合可以提高数据识别的准确率.Khammassi 等人[5]采用遗传算法和逻辑回归算法进行特征选择,选取最优的特征子集.通过不同的决策树算法证明本方法选取的特征子集对于入侵检测是有效的.Aljawarneh 等人[6]采用信息增益来选取重要特征,并提出一种混合投票模型,结合 J48,Meta Paging,Random Tree 等方法来识别异常数据,该方法能够提高检测的准确性,降低误报率.George[7]选择支持向量机和主成分分析来执行网络数据的异常检测,与贝叶斯分类算法和信息增益降维方法的分类效果进行比较,并证明主成分分析和信息增益降维方法都能改善贝叶斯分类算法的效果,但会降低支持向量机的分类效果[8].

除了上述方法外,一些方法还侧重于研究数据集中数据项的选取,通过采用聚类方法或抽样方法来选取部分有代表性的训练数据.该训练数据与机器学习算法相结合,可以显著降低分类器的训练时间并提高分类的准确率.程晓旭等人[9]采用改进的 K-means 算法得到全局的最优的聚类划分,降低了异常检测的时间复杂度.Alyaseen 等人[10-12]将改进的 K-means 方法与机器学习方法相结合来构建入侵检测模型.改进的 K-means 方法能发现数据之间相似的结构或模型,从而能够降低数据集的长度,得到一个高质量的数据集.将改进的 K-means 与 C4.5 结合来构造入侵检测模型的分类器,大大降低了入侵检测系统的运行时间[10];与支持向量机算法相结合有效的提高了异常数据类型 DoS(denial of service),R2L(remote to local),U2R(user to root)的检测率[11];与支持向量机和极限学习机(extreme learning machine, ELM)的混合模型相结合来提高入侵检测系统的准确性和效率[12].Roshan 等人[13]提出一种自适应的入侵检测系统结合聚类和极限学习机模型,该方法能够以较小的代价检测出已知的和新型的攻击.Enamul 等人[14]采用抽样技术与最小二乘支持向量机(least square support vector machine, LS-SVM)的混合模型,抽样技术选择最有代表性的训练数据集,最小二乘支持向量机对网络数据进行分类,标识异常数据类型.

尽管对于网络异常行为检测的研究很多,但仍然存在着某些异常行为(例如 Probe(probing),U2R,

R2L)的检测率较低以及各类别检测不均衡的问题.本文主要针对这一问题,提出一个新的混合模型,结合 KNN(K nearest neighbors)离群点检测算法和多层随机森林算法来建立入侵检测模型.采用 KNN 离群点检测算法获取一个小规模、高质量的训练数据集.在此训练数据集上,结合类别检测划分方法,构建多层次的随机森林模型,检测网络异常行为.基准数据集 KDD(knowledge discovery and data mining) Cup 1999 用来评估本文算法的正确性和检测性能,并在不同大小的测试数据集上验证本文算法的扩展性.

## 1 基于 KNN 离群点检测和随机森林的多层入侵检测方法

本节描述提出的方法结合数据选取和多层随机森林算法来进行入侵检测.该方法主要包括 4 个阶段:1)对训练数据集和测试数据集进行预处理;2)数据选取阶段,得到一个新的规模小、质量高的训练数据集;3)采用新的训练数据集训练多层的随机森林分类器;4)测试校准测试数据集.最终得到测试数据集中正常行为和异常行为的检测结果.本文算法的整体流程如图 1 所示:

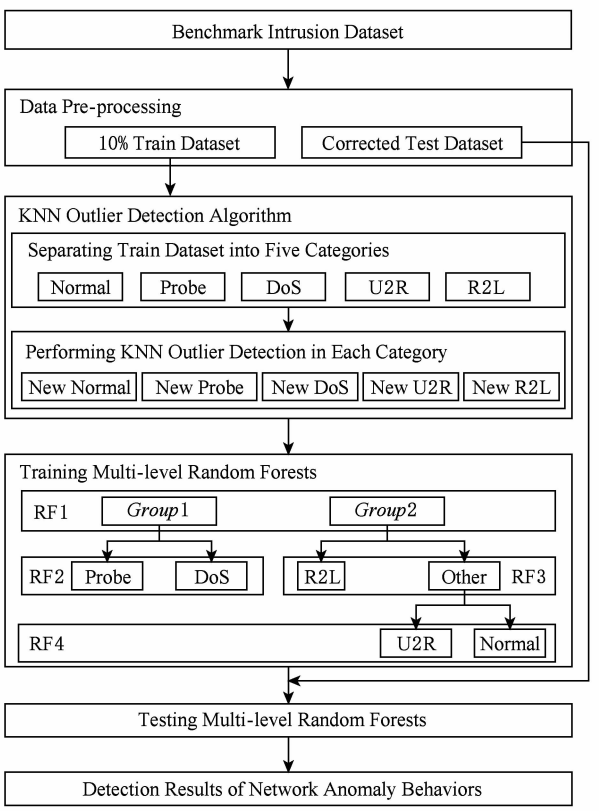



Fig. 1 The entire flow of the proposed method

图 1 本文算法的整体流程图

在基准数据集 KDD Cup 1999 中,10%的训练数据集具有数据量大和数据类别不平衡的问题,该数据集被完全用于分类器时,存在许多引人注意的问题.首先是训练器的训练时间会非常长,可能会因为内存溢出而导致训练失败.其次,因为数据之间的不平衡,导致训练器在分类性能上更偏向于数量较多的类别,使得类别 Probe,U2R,R2L 的分类准确性偏低.因此本文对训练数据集进行数据抽取,从而降低数据集的大小,并得到新的高质量的训练数据集.

### 1.1 基于 KNN 离群点检测的数据选择方法

该部分采用 KNN 离群点检测算法对数据进行选择.KNN 离群点检测算法是一种基于距离的离群点检测算法,最早由 Knorr 等人在 1998 年提出.该方法是在 KNN 的基础上发展而来的,是一种比较简单且易于应用的离群点检测算法.KNN 离群点检测算法的基本思想是通过计算数据集 $D$ 中每个数据与数据集 $D$ 中其他数据的 $K$ 近邻平均距离,并对每个点的 $K$ 近邻平均距离进行降序排序,距离最大的前 $N$ 个点被认为是离群点.在本文中,离群点被认为是分布稀疏且离高密度的群体较远的点.在数据选择时,删除数据集 $D$ 中的 $N$ 个离群点,得到新的数据集 $D'$,新数据集的大小为 $M=|L-N|$,$L$ 为原数据集 $D$ 的大小.

在采用 KNN 离群点检测算法进行数据选择之前,先将训练数据集分为 5 类:Normal,Probe,DoS,U2R,R2L,在每一类数据集上执行 KNN 离群点检测算法,得到一个新的规模小、质量高的训练数据集,该数据集可以有效地改善多层随机森林分类器的训练时间和性能.在网络流量中,由于攻击行为 U2R 和 R2L 的样本数量的所占比例非常少,在分类效果上处于劣势,因此 U2R 和 R2L 类型不执行离群点删除操作.在 Normal,Probe,DoS 这 3 个类别上执行离群点检测算法,每个类别的离群点检测算法设置不同的参数 $M$ 和 $K$.在本文中通过多次实验获得每个类别中最优的参数 $M$ 和 $K$,从而得到最优的网络异常行为的检测率.算法 1 展示了根据 KNN 离群点检测算法选取新的训练数据集的详细步骤.

**算法 1.** 基于 KNN 离群点检测的数据选择算法.

输入:10%的训练数据集 $TD$、参数 $K_{\text{Normal}}$,$M_{\text{Normal}}$,$K_{\text{Probe}}$,$M_{\text{Probe}}$,$K_{\text{DoS}}$,$M_{\text{DoS}}$;

输出:新的数据集 $D'_{\text{Normal}}$,$D'_{\text{Probe}}$,$D'_{\text{DoS}}$.

① $D_{\text{Normal}}$,$D_{\text{Probe}}$,$D_{\text{DoS}}=\varnothing$; /* 存储不同类别的数据集合 */

② for each $d$ in $TD$

③ if ($d.label==1$) then $D_{\text{Normal}}.add(d)$;
/* 数据 $d$ 的 $label$ 为 1,表示该数据的类别为 Normal;$label$ 为 2,表示该数据是 Probe 类别;$label$ 为 3,表示该数据是 DoS 类型 */

④ else if ($d.label==2$) then $D_{\text{Probe}}.add(d)$;

⑤ else ($d.label==3$) then $D_{\text{Dos}}.add(d)$;

⑥ end if

⑦ end for

⑧ $D'_{\text{Normal}}=KOD(D_{\text{Normal}},K_{\text{Normal}},M_{\text{Normal}})$;

⑨ $D'_{\text{Probe}}=KOD(D_{\text{Probe}},K_{\text{Probe}},M_{\text{Probe}})$;

⑩ $D'_{\text{DoS}}=KOD(D_{\text{DoS}},K_{\text{DoS}},M_{\text{DoS}})$.

produce $KOD(D,K,M)$

① $L=D.length$; /* 数据集 $D$ 的大小 */

② $Avg[L]$, $Index=\varnothing$; /* $Avg$ 中存储每条数据的 $K$ 近邻平均距离,$Index$ 中存储前 $M$ 个数据的索引值 */

③ $d[L]=\varnothing$; /* 存储数据之间的欧氏距离 */

④ for ($i=0;i<L;i++$)

⑤ for ($j=0;j<L;j++$)

⑥ 计算数据 $D(i)$ 和 $D(j)$ 之间的欧氏距离 $d[j]$;

⑦ end for

⑧ 对 $d$ 进行升序排序;

⑨ 计算 $d$ 中的前 $K$ 个数据的平均值 $Avg[i]$;

⑩ end for

⑪ 对 $Avg$ 进行升序排序;

⑫ 得到前 $M$ 个数据的索引 $Index$;

⑬ 根据 $Index$ 得到新的数据集 $D'$.

### 1.2 多层随机森林模型

随机森林(random forests, RF)最早是由 Leo[15] 作为一个分类算法提出的,广泛应用于入侵检测和计算生物学等方面.该算法的优势在于:该算法是一种集成学习方法,对于任何类型的数据集,随机森林算法的分类效果和聚类效果要优于大多数算法,并且能够有效处理高维度的数据集.该算法对于参数的设置并不敏感,可以很容易地找到一个合适的随机森林模型.随机森林算法是一个组合分类器,以决策树为基础分类器.该模型的基本思想是:一个森林包含多个决策树,每棵决策树是由有放回的随机抽样样本构造的,也就是说在总的训练集中的有些样本可能多次出现在 1 棵树的训练集中,也可能从未出现在 1 棵树的训练集中,并使每棵决策树进行充分生长,不进行任何剪枝,最终的输出结果就是所有决策树进行多数投票的结果.

本文提出以随机森林模型作为基础分类器来构建多层的异常检测分类器.为了更有效地检测异常行为,提出一种新的类别检测划分方法,该方法与多个的随机森林分类器相结合来构建多层的异常检测分类器.新的划分方法和多层随机模型结构如图 2 所示.该模型包含 4 个随机森林分类器,第 1 个分类器(RF1)将数据分为 2 组:$Group1$ 和 $Group2$,$Group1$ 包括 DoS 和 Probe,$Group2$ 包括 Normal,U2R,R2L;第 2 个分类器(RF2)区分 DoS 和 Probe;第 3 个分类器(RF3)检测 R2L;第 4 个分类器(RF4)是检测 U2R 和 Normal.根据经验将随机森林模型中树的数量设置为 150.

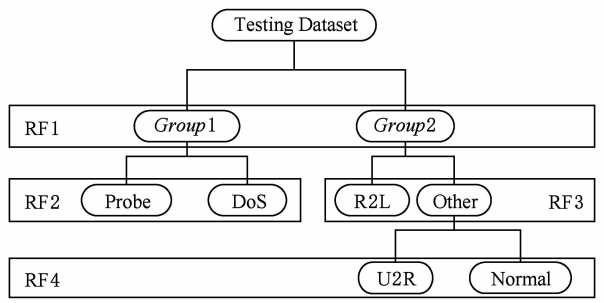



Fig. 2 Multi-level random forests model

图 2 多层随机森林模型

根据网络流量的相似性,本文提出一种新的类别检测划分方法.该划分方法首先将网络数据划分为 2 组.对于第 1 组来说,DoS 和 Probe 在流量特征上更相似,同时与其他的类别有更少的相似性,因此将这 2 个类别的数据分为 1 组.对于第 2 组来说,当发生 U2R 和 R2L 攻击时,此时的流量特征与正常的连接区别很小,Normal,U2R,R2L 在流量特征上更相似[16],因此将这 3 种类别分为 1 组.这种划分尽可能避免了异常行为在检测时的相互干扰,尤其是异常行为 Probe,U2R,R2L 的检测.在网络数据中,攻击类型的数目要远小于正常连接的数目,其中 Probe,U2R,R2L 类别的数据所占的比例非常小,这就使得异常行为检测算法需要在检测率与误报率之间寻找一个平衡,尽可能多得检测出攻击行为,同时避免将正常行为误检为攻击行为.DoS 属于大流量的攻击行为,检测相对更容易一些.U2R 和 R2L 属于明显的小流量攻击,并且造成的危害更大,其中 U2R 被认为是网络中最危险的行为[17].Probe 是对网络的扫描和检测,被认为是多种严重攻击的前提行为.因此对 Probe,U2R,R2L 的检测是至关重要

的.与其他算法相比,本文提出的划分方法和多层的随机森林模型能有效地检测出 Probe,U2R,R2L,并在所有类别的检测间取得一个平衡,因此,本文提出的划分法和检测模型是合理和有效的.

## 2 实验结果

本节评估本文提出的算法的性能,所有的实验均在 Windows 7 PC,Inter® Pentium® CPU G2020 @2.90 GHz,4.00 GB RAM 环境中实现.采用 MATLAB 7.8.0 实现本文的算法.

### 2.1 训练集与测试集

KDD Cup 1999 数据集是入侵检测领域的基准数据集,被广泛用于入侵检测系统的研究中.该数据集一共包括 41 个属性和 1 个类别标签(正常或其他的攻击类型),其中 7 个属性(属性 2,3,4,7,12,14,22)是离散属性,其他属性是连续属性.在离散属性中属性 2,3,4 是符号属性,其他是数值属性.这 41 个属性可以分为 4 类:TCP 连接基本特征、TCP 连接内容特征、基于时间的网络流量统计特征和基于主机的网络流量统计特征,如表 1 所示.数据集中异常数据类型分为 4 类:DoS,Probe,U2R,R2L.KDD Cup 1999 数据集提供了训练和测试数据集.10% KDD 训练数据集包括 22 种攻击类型,测试数据集中还包括额外的 17 种攻击类型.KDD Cup 1999 训练数据集和测试数据集的详细信息如表 2 所示.

**Table 1 Attributes of KDD Cup 1999 Dataset**

**表 1 KDD Cup 1999 数据集的属性**

| Class | Attribute Name |
|---|---|
| Basic Features | duration,protocol_type,service,flag,src_bytes,dst_bytes,land,wrong_fragment,urgent |
| Content Features | hot,num_failed_logins,logged_in,num_compromised,root_shell,su_attempted,num_root,num_file_creations,num_shells,num_access_files,num_outbound_cmds,is_hot_login,is_guest_login |
| Time Based Traffic Features | count,srv_count,serror_rate,srv_serror_rate,rerror_rate,srv_rerror_rate,same_srv_rate,diff_srv_rate,srv_diff_host_rate |
| Host Based Traffic Features | Dst_host_count,dst_host_srv_count,dst_host_same_srv_rate,dst_host_diff_srv_rate,dst_host_same_src_port_rate,dst_host_srv_diff_host_rate,dst_host_serror_rate,dst_host_srv_serror_rate,dst_host_rerror_rate,dst_host_srv_rerror_rate |

**Table 2 Details of Training Dataset and Testing Dataset in KDD Cup 1999**

**表 2 KDD Cup 1999 训练数据集和测试数据集中不同类别数据的详细信息**

| Category | 10% KDD Dataset Training Dataset | | Corrected Dataset Testing Dataset | | | |
|---|---|---|---|---|---|---|
| | Category | Number of Each Category in Training Dataset | Known Category | Number of Each Known Category in Testing Dataset | Unknown Category | Number of Each Unknown Category in Testing Dataset |
| Normal | normal | 97 278 | Normal | 60 593 | | |
| Probe | Ipsweep | 1 247 | Ipsweep | 306 | Mscan | 1 053 |
| | Nmap | 231 | Nmap | 84 | Saint | 736 |
| | Portsweep | 1 040 | Portsweep | 354 | | |
| | Satan | 1 589 | Satan | 1 633 | | |
| DoS | Back | 2 203 | Back | 1 098 | Apache2 | 794 |
| | Land | 21 | Land | 9 | Mailbomb | 5 000 |
| | Neptune | 107 201 | Neptune | 58 001 | Processtable | 759 |
| | Pod | 264 | Pod | 87 | Udpstorm | 2 |
| | Smurf | 280 790 | Smurf | 164 091 | | |
| | Teardrop | 979 | Teardrop | 12 | | |
| U2R | Buffer_overflow | 30 | Buffer_overflow | 22 | Httptunnel | 158 |
| | Loadmodule | 9 | Loadmodule | 2 | Ps | 16 |
| | Perl | 3 | Perl | 2 | Sqlattack | 2 |
| | Rootkit | 10 | Rootkit | 13 | Xterm | 13 |

**Continued (Table 2)**

| Category | 10% KDD Dataset Training Dataset | | Corrected Dataset Testing Dataset | | | |
|---|---|---|---|---|---|---|
| | Category | Number of Each Category in Training Dataset | Known Category | Number of Each Known Category in Testing Dataset | Unknown Category | Number of Each Unknown Category in Testing Dataset |
| R2L | Ftp_write | 8 | Ftp_write | 3 | Named | 17 |
| | Guess_passwd | 53 | Guess_passwd | 4 367 | Sendmail | 17 |
| | Imap | 12 | Imap | 1 | Worm | 2 |
| | Multihop | 7 | Multihop | 18 | Xlock | 9 |
| | Phf | 4 | Phf | 2 | Xsnoop | 4 |
| | Spy | 2 | Warezmaster | 1 602 | Snmpgetattack | 7 741 |
| | Warezclient | 1 020 | | | Snmpguess | 2 406 |
| | Warezmaster | 20 | | | | |

10%的训练数据集和校准测试数据集在应用到入侵检测方法之前,首先需要进行数据预处理.数据预处理过程共包括 5 个步骤:1)由于训练数据集中包含重复的数据,首先对训练数据集进行去重处理,将训练数据集由 494 021 条数据降低为 145 586 条数据.2)由于训练数据集中属性列 num_outbound_cmds,is_hot_login 的所有数值均为 0,对数据的分类没有任何影响,因此将训练数据集和测试数据集中的属性列 num_outbound_cmds,is_hot_login 删除.3)将训练集和测试集中的符号属性 protocol_type,service,flag 转化为数值属性.以属性 protocol_type 为例,该属性共包括 3 种:TCP,UDP,ICMP,将这 3 种类别用数值表示,即 1 表示 TCP,2 表示 UDP 以及 3 表示 ICMP.4)将类别标签转化为数值表示,其中 1 表示 Normal 类别,2 表示 Probe 类别,3 表示 DoS 类别,4 表示 U2R 类别以及 5 表示 R2L 类别.5)将训练数据集和测试数据集进行[0,1]标准化处理.采用 min-max 标准化方法对训练集和测试集进行标准化[12]:

$$v'=\frac{v-\min\limits_{i}}{\max\limits_{i}-\min\limits_{i}}, \tag{1}$$

其中,$v$ 是第 $i$ 个属性列的一个值,$\min\limits_{i}$是第 $i$ 个属性列的最小值,$\max\limits_{i}$是第 $i$ 个属性列的最大值.

## 2.2 实验评估指标

入侵检测系统中存在许多可利用的评估指标.其中准确性 $Acc$(accuracy)、检测率 $DR$(detection rate)和误报率 $FAR$(false alarm rate)是入侵检测系统中主要的评估指标[4].

$$Acc=\frac{TP+TN}{TP+TN+FP+FN}, \tag{2}$$

$$DR=\frac{TP}{TP+FN}, \tag{3}$$

$$FAR=\frac{FP}{TN+FP}, \tag{4}$$

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.

## 2.3 实验分析

第 1 个实验用来评估 KNN 离群点算法检测的性能.KNN 离群点检测算法的目标是用来提高多层随机模型分类器的性能.该实验通过比较 KNN 离群点检测算法和多层随机森林算法构成的混合模型与单一的多层随机森林模型的检测效果,来说明 KNN 离群点检测算法能改善分类器的性能.为了这一目的,首先选取 KNN 的离群点检测算法中最优的参数 $K$ 和 $M$.参数的选取分为 3 个部分,Probe 类别的参数、DoS 类别的参数和 Normal 类别的参数.首先选取 Probe 类别的参数,将 $K_{\text{Probe}}$ 分为 5 个层次 10,20,30,40,50,在每个层次下选择不同的 $M_{\text{Probe}}$来进行实验,实验结果显示在表 3 中.当 $K_{\text{Probe}}$

**Table 3 Detection Rate of Probe in Different Parameters**

**表 3 Probe 在不同的参数下的检测率**

| $M$ | $K$ | | | | |
|---|---|---|---|---|---|
| | 10 | 20 | 30 | 40 | 50 |
| 1 500 | 0.872 7 | 0.867 4 | 0.879 7 | 0.892 4 | 0.887 7 |
| 1 600 | 0.879 9 | 0.886 5 | 0.900 3 | 0.885 1 | 0.903 4 |
| 1 700 | 0.869 6 | 0.894 4 | 0.906 9 | 0.894 2 | 0.909 1 |
| 1 800 | 0.906 1 | 0.901 2 | 0.908 5 | 0.912 9 | 0.930 9 |
| 1 900 | 0.924 8 | 0.933 6 | 0.945 9 | 0.943 5 | 0.948 6 |
| 2 000 | 0.940 2 | 0.952 3 | 0.954 6 | 0.954 7 | 0.954 7 |

值一定，$M_{Probe}$ 增加时，Probe 的检测率 $DR_{Probe}$ 不断上升. 当 $M_{Probe}$ 一定，$K_{Probe}$ 增加时，Probe 的检测率有稍微的增加. 根据表 3 分析，当 $M_{Probe}=2\,000$ 时，检测率是最高的，在 $K_{Probe}=40$ 和 50，Probe 类别的检测率相同. 随着 $K_{Probe}$ 不断增大，Probe 的检测率基本上保持不变. 因此 Probe 的参数设置为 $K_{Probe}=40$，$M_{Probe}=2\,000$.

采用与 Probe 参数选择类似的方法，通过多次实验，选择 DoS 和 Normal 类别的参数. 对于 DoS 类别来说，DoS 类别的数据量将会影响 RF2 分类器检测 DoS 的检测率，因此由 DoS 的检测率 $DR_{DoS}$ 来决定 $K_{DoS}$ 和 $M_{DoS}$ 的选取. $K_{DoS}$ 从 5，10，20，30 共 4 个层次进行分析，$M_{DoS}$ 从 10 000，15 000，20 000 共 3 个层次进行分析，实验结果显示在 $K_{DoS}=10$，$M_{DoS}=15\,000$ 时 DoS 的检测率最高. 为了得到更好的 $M_{DoS}$ 的取值，将 $M_{DoS}$ 从 10 000～15 000 进行细粒度的划分. 根据实验结果分析，将 DoS 的参数设置为 $K_{DoS}=10$，$M_{DoS}=11\,000$. 对于 Normal 类别来说，Normal 类别的数据量影响 Normal 和 R2L 的检测率，因此根据 Normal 和 R2L 的检测率来决定 $K_{Normal}$ 和 $M_{Normal}$ 的选取. 在实验中将 $K_{Normal}$ 的取值设置为 50，100，150，200，250，300 共 6 个层次，$M_{Normal}$ 依次设置为 2 000，3 000，4 000，5 000，6 000，7 000，8 000，9 000 共 8 个层次进行分析，实验结果显示：随着 $K_{Normal}$ 的增大，Normal 类别的检测率 $DR_{Normal}$ 先增大后减小；随着 $M_{Normal}$ 的增大，Normal 的检测率不断增大，但变化的幅度并不大. 但随着 $M_{Normal}$ 和 $K_{Normal}$ 的不断增加，R2L 的识别准确率逐渐降低. 为使得 Normal 和 R2L 都有较高的检测率，本文选取 $M_{Normal}=6\,000$，$K_{Normal}=150$.

通过上述实验分析，在 Normal，Probe，DoS 这 3 个类别下，KNN 离群点检测算法的参数 $M$ 和 $K$ 设置如表 4 所示. 应用 KNN 离群点检测算法之前和之后不同类别的数据集的数量如表 5 所示.

**Table 4 Parameters of *KNN* Outlier Detection Algorithm**

**表 4 *KNN* 离群点检测算法的参数设置**

| Category | $K$ | $M$ |
|---|---|---|
| Normal | 150 | 6 000 |
| Probe | 40 | 2 000 |
| DoS | 10 | 11 000 |

**Table 5 Size of Dataset Before and After Applying *KNN* Outlier Detection Algorithm**

**表 5 应用 *KNN* 离群点检测算法之前和之后不同种类数据的大小**

| Category | Detection Before | Detection After |
|---|---|---|
| Normal | 87 832 | 6 000 |
| Probe | 2 131 | 2 000 |
| DoS | 54 572 | 11 000 |
| U2R | 52 | 52 |
| R2L | 999 | 999 |

新的小规模、高质量的数据集用于训练多层的随机森林分类器. 图 3 展示了混合的 KNN 离群点检测算法和多层随机森林的模型(混合模型)与单一的多层随机森林模型(单一模型)的检测性能，混合模型在准确率和检测率上要优于单一模型，并有一个可以接受的误报率. 对异常类型 Probe，U2R，R2L 的检测效果明显优于单一模型，DoS 的检测效果略高于单一模型，Normal 类型的检测率略微低于单一的模型，但仍然是很好的识别结果. 该实验说明 KNN 离群点检测算法可以有效的提高多层随机森林模型的准确率和检测率，并能提高 DoS，Probe，U2R，R2L 异常类型的检测率.

为了说明多层随机森林算法的性能要优于其他

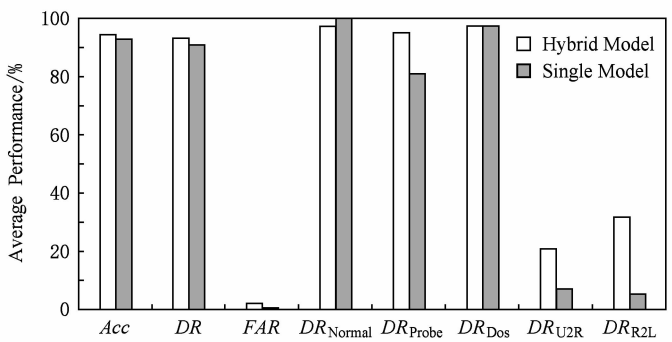



Fig. 3 The performance comparison between the hybrid model and the single model

图 3 混合模型和单一模型的性能比较

分类算法这一问题,本文将混合的多层极限学习机算法(混合 ELM 模型)与混合的多层随机森林算法(混合 RF 模型)进行对比,图 4 说明了多层随机森林算法的准确率、检测率以及对各个数据类的检测率都要优于多层极限学习机模型,同时该模型具有较低的误报率.

为了说明本文算法的扩展性,将本文提出的算法在不同大小的测试集上进行测试,检测其准确率、检测率、误报率以及不同类别的检测率是否随着数据集的变化保持稳定.从校准测试数据集中随机无放回抽取数据,构成 3 个不同大小的测试数据集(25%校准数据集,50%校准数据集和 100%校准数据集),如表 6 所示.图 5 展示了不同测试集的检测效果,随着测试集的增大,本文提出的算法在准确率、检测率和误报率上均能保持稳定.对于不同类别的检测率,除了 DoS 的检测率下降了大约 3%,其他类别的检测率均保持稳定或略微提高.说明本文提出的算法能适应不同大小的测试数据集,有很好的扩展性.

**Table 6 Different Size of Three Testing Dataset**

**表 6 3 个不同的测试数据集的数据量**

| Category | 25% Corrected Dataset | 50% Corrected Dataset | 100% Corrected Dataset |
|---|---|---|---|
| Normal | 20 296 | 40 372 | 80 889 |
| Probe | 1 363 | 2 784 | 5 529 |
| DoS | 76 585 | 153 246 | 306 437 |
| U2R | 76 | 156 | 304 |
| R2L | 5 356 | 10 795 | 21 545 |
| Total | 103 676 | 207 353 | 414 704 |

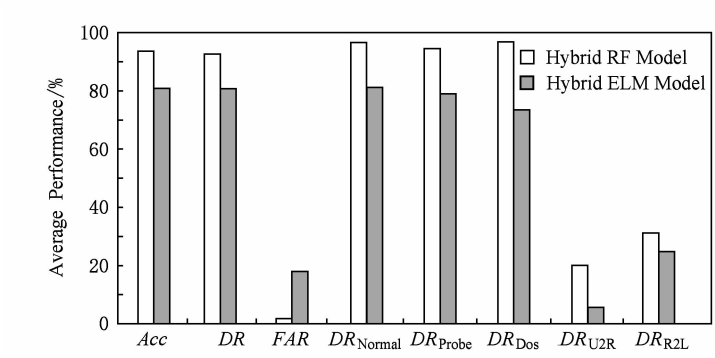



Fig. 4 The performance comparison between the hybrid RF model and the hybrid ELM model

图 4 混合的随机森林模型和极限学习机模型的性能比较

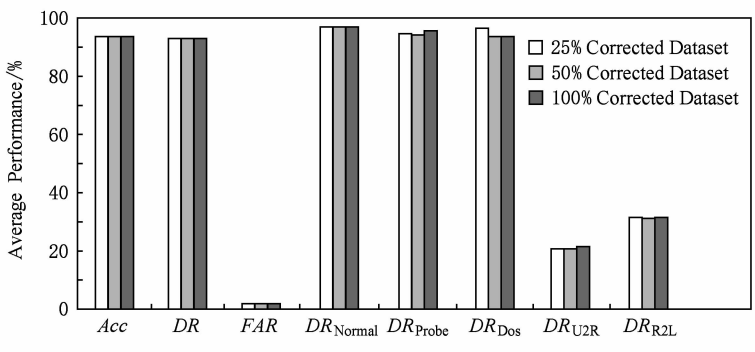



Fig. 5 The detection performance of different testing dataset

图 5 不同测试集的检测性能

入侵检测方法在网络异常检测时要求做到网络异常行为的实时性检测.为了说明多层随机森林算法检测的检测效率,将多层随机森林算法与单一随机森林算法的检测时间在多个不同规模的测试集上

进行对比,如图 6 所示.本文提出的多层随机森林算法比单一随机森林算法的运行时间要稍微高一点,但在不同规模的数据集上,运行时间仅差 1 s 左右.在整个测试数据集上,检测时间不到 8 s,因此本文提出的方法可以高效地检测网络入侵行为,满足入侵检测中实时性的要求.

为了更好地验证本文提出的算法,在整个校准数据集上将本文算法与其他 4 种算法相比较,表 7 展示了比较的结果.表 7 说明本文提出的算法具有较高的准确率和检测率,同时有一个可以接受的误报率.在 Probe,U2R,R2L 的检测上要明显优于其他算法,Normal,DoS 的检测率略微低于其他算法.传统的极限学习机算法和随机森林算法能更好的识别出正常的行为,对于异常行为,尤其是 Probe,U2R,R2L 的检测率均比较低.Genetic 算法有较高的误报率,这是因为正常的行为被大量误判为异常的行为.Multiclass SVM 同样对于 Probe,U2R,R2L 的检测率较低.本文算法的优势在于有较高的准确率和检测率,能很好的检测出 Probe,U2R,R2L,同时在每个类别的检测率之间取得一个平衡.

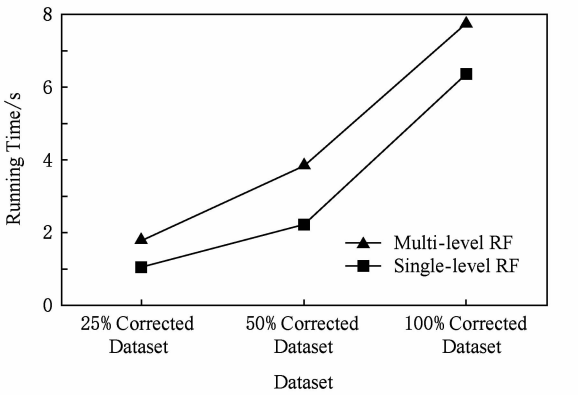



Fig. 6 The running time comparison between multi-level RF and the single-level RF

图 6 多层和单层随机森林运行时间比较

**Table 7 Performance Comparison Between the Proposed Method and Other Algorithms**

**表 7 本文算法和其他算法的性能比较**

| Comparison Index | Proposed Method | ELM | Random Forests | Genetic Algorithm[18] | Multiclass SVM[19] |
|---|---|---|---|---|---|
| *Acc* | **0.943 6** | 0.928 7 | 0.926 6 | 0.90 | 0.924 6 |
| *DR* | 0.935 5 | 0.912 8 | 0.909 6 | **0.9495** | 0.907 4 |
| *FAR* | 0.023 4 | 0.016 5 | 0.009 9 | 0.3046 | **0.004 3** |
| Normal | 0.976 6 | 0.993 6 | **0.996 2** | 0.695 | 0.996 |
| Probe | **0.953 4** | 0.755 6 | 0.757 8 | 0.711 | 0.75 |
| DoS | 0.973 2 | 0.972 8 | 0.972 5 | **0.994** | 0.968 |
| U2R | **0.210 5** | 0.026 3 | 0.078 9 | 0.189 | 0.053 |
| R2L | **0.319 6** | 0.074 4 | 0.039 7 | 0.054 | 0.042 |

Note: The font of bold type indicates the best detection result achieved by the proposed or comparied methods.

## 3 总 结

在大量的网络数据中,为了更有效地检测网络异常行为,本文提出一种新的混合的入侵检测方法,该方法结合 *K*NN 离群点检测算法和多层次的随机森林模型来检测异常的网络行为.采用 *K*NN 离群点检测算法来检测并删除训练数据集中的离群数据,得到一个小规模、高质量的训练数据集,该数据集可以有效地改善分类器的训练时间和检测性能.根据网络流量的相似性,提出一种新的类别检测划分方法,该方法能对不同的类别进行有效的辨别.结合这种划分方法,提出的多层随机森林模型能有效的检测异常行为.基准数据集 KDD Cup 1999 用来评估本文算法的性能.不同大小的测试数据集用来评估本文算法,说明该算法具有良好的扩展性.与其他算法相比,本文的算法准确率要明显优于其他算法.该算法的主要贡献在于对 Probe,U2R,R2L 这 3 个异常类别的检测,其检测率要远远优于其他算法,并在不同类别的检测率之间取得一个平衡.

## 参 考 文 献

[1] Lee W, Stolfo S J, Mok K W. A data mining framework for building intrusion detection models [C] //Proc of the 20th IEEE Symp on Security & Privacy. Piscataway, NJ: IEEE, 1999: 120-132

[2] Roesch M. Snort-lightweight intrusion detection for networks [C] //Proc of the 13th USENIX Conf on System Administration. Berkeley, CA: USENIX Association, 1999: 229-238

[3] Om H, Kundu A. A hybrid system for reducing the false alarm rate of anomaly intrusion detection system [C] //Proc of the 1st Int Conf on Recent Advances in Information Technology. Piscataway, NJ: IEEE, 2012: 131-136

[4] Raman M R G, Somu N, Kirthivasan K, et al. An efficient intrusion detection system based on hypergraph-genetic algorithm for parameter optimization and feature selection in support vector machine [J]. Knowledge-Based Systems, 2017, 134: 1-12

[5] Khammassi C, Krichen S. A GA-LR wrapper approach for feature selection in network intrusion detection [J]. Computers & Security, 2017, 70: 255-277

[6] Aljawarneh S, Aldwairi M, Yassein M B. Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model [J]. Journal of Computational Science, 2018, 25: 152-160

[7] George A. Anomaly detection based on machine learning dimensionality reduction using PCA and classification using SVM [J]. International Journal of Computer Applications, 2012, 47(21): 5-8

[8] Hashem S H. Efficiency of SVM and PCA to enhance intrusion detection system [J]. Journal of Asian Scientific Research, 2013, 3(4): 381-395

[9] Cheng Xiaoxu, Yu Haitao, Li Zi. Improved *K*-means network intrusion detection algorithm [J]. Intelligent Computer & Applications, 2012, 2(2): 21-23 (in Chinese) (程晓旭, 于海涛, 李梓. 改进的 *K*-means 网络入侵检测算法[J]. 智能计算机与应用, 2012, 2(2): 21-23)

[10] Alyaseen W L, Othman Z A, Nazri M Z A. Hybrid modified *K*-means with C4.5 for intrusion detection systems in multiagent systems [J]. The Scientific World Journal, 2015, 2015(2): 294761

[11] Alyaseen W L, Othman Z A, Nazri M Z A. Intrusion detection system based on modified *K*-means and multi-level support vector machines [C] //Proc of the 1st Int Conf on Soft Computing in Data Science. Berlin: Springer, 2015: 265-274

[12] Alyaseen W L, Othman Z A, Nazri M Z A. Multi-level hybrid support vector machine and extreme learning machine based on modified *K*-means for intrusion detection system [J]. Expert Systems with Applications, 2017, 67: 296-303

[13] Roshan S, Miche Y, Akusok A, et al. Adaptive and online network intrusion detection system using clustering and extreme learning machines [J]. Journal of the Franklin Institute, 2018, 355(4): 1752-1779

[14] Enamul K, Hu Jiankun, Wang Hua, et al. A novel statistical technique for intrusion detection systems [J/OL]. Future Generation Computer Systems. 2017 [2017-11-08]. https://www.researchgate.net/publication/313034553_A_novel_statistical_technique_for_intrusion_detection_systems

[15] Leo B. Random forests [J]. Machine Learning, 2001, 45(1): 5-32

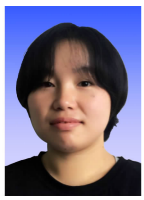
[16] Gogoi P, Bhattacharyya D K, Borah B, et al. MLH-IDS: A multi-level hybrid intrusion detection method [J]. Computer Journal, 2014, 57(4): 602-623

[17] Guo Yutong, Wang Yan, Qin Mengyuan, et al. DPI & DFI: A malicious behavior detection method combining deep packet inspection and deep flow inspection [J]. Procedia Engineering, 2017, 174: 1309-1314

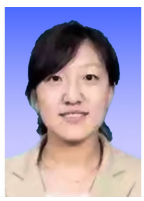
[18] Hoque M S, Mukit M A, Bikas M A N. An implementation of intrusion detection system using genetic algorithm [J]. International Journal of Network Security & Its Applications, 2012, 4(2): 109-120

[19] Ambwani T. Multi class support vector machine implementation to intrusion detection [C] //Proc of the 14th Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2003: 2300-2305

**Ren Jiadong**, born in 1967. PhD, professor, and member of CCF. His main research interests include data mining and software security.



**Liu Xinqian**, born in 1992. PhD candidate. Her main research interests include network security, and social network. (1689890718@qq.com)



**Wang Qian**, born in 1987. PhD, lecturer. Her main research interests include data mining, software security and network security. (wangqianysu@163.com)

**He Haitao**, born in 1968. PhD, professor. Her main research interests include data mining and software security. (haitao@ysu.edu.cn)

**Zhao Xiaolin**, born in 1971. PhD, associate professor. His main research interests include complex network, software security, network security. (zhaoxl@bit.edu.cn)