一种基于FRS-FCM算法的集成入侵检测方法的研究

刘永忠 李欣娣 李 杨 张为群

(西南大学计算机与信息科学学院 重庆 400715) (重庆市智能软件与软件工程重点实验室 重庆 400715)

摘 要 传统FCM算法对初值的依赖性过大且欧氏距离只适用于处理数值型及特征空间为超球结构的数据集。为此,利用模糊粗糙集思想,结合ReliefF技术,提出了一种基于模糊粗糙集的特征加权聚类算法(FRS-FCM),并将此算法应用到集成入侵检测中,通过有效地聚类和集成学习来提高入侵检测的检测率,降低误检率,并较大地提高低频攻击的检测率。最后利用KDD Cup 99数据集进行的仿真实验验证了该方法的可行性与有效性。

关键词 模糊粗糙集,ReliefF,FRS-FCM,集成入侵检测

Research Based on a Method of FRS-FCM Ensemble Intrusion Detection

LI Yong-zhong LI Xing-di LI Yang ZHANG Wei-qun

(Faculty of Computer and Information Science, Southwest University, Chongqing 400715, China)

(Chongqing Intelligent Software and Software Engineering Laboratory, Chongqing 400715, China)

Abstract Traditional FCM algorithm is too dependent on initial distance and Euclidean distance is only applied to handle the dataset of numeric and spatial data structure for the super-ball. Based on fuzzy rough sets and ReliefF technology, the author proposed a fuzzy rough set based clustering algorithm(FRS-FCM), and used it to integrated intrusion detection. By effective clustering and integrated learning, the algorithm can improve the detection rate and reduce the false detection rate, improve the detection rate of low-frequency attacks. Finally, simulation experiments using KDD Cup 99 data set verify feasibility and effectiveness of the algorithm.

Keywords Fuzzy rough sets, ReliefF, FRS-FCM, Integrated intrusion detection

1 引言

入侵检测通过对主机或网络中的事件进行分析、比较和学习,来发现对计算机造成安全威胁的入侵行为,进而触发相应的响应措施。国内外研究学者在改善入侵检测性能方面做了大量的研究:国外学者先后将贝叶斯网络、决策树、数据挖掘、神经网络、遗传算法等机器学习理论[1]运用到入侵检测中,从而在一定程序上提高了检测速度、自学习与自适应性,并降低了误报与漏报率;国内学者则倾向于融合算法[2,8]的研究,使各算法互相弥补不足,共同提高入侵检测的性能。

但传统的IDS一般都存在着误报与漏报率高、自适应性差、检测分析方法单一等不足。因此本文提出一种基于FRS-FCM算法的集成入侵检测方法,它克服了传统FCM算法对初始值依赖性大、只能处理数值型数据和只适用于特征空间为超球结构的数据集等缺点,在提高入侵检测速度与自适应性的同时,降低了误报率与漏报率,并对低频攻击的检测率有较大的提高。

2 基本概念

2.1 相关定义

定义1(模糊上下近似) 知识表达系统 $K=(U,R)$, $U=\{x_1,\cdots,x_n\}$为论域,R是论域U上的一个等价关系,$F=\{F_1,F_2,\cdots,F_i,\cdots,F_k\}i\in\{1,2,\cdots,k\}$为$R$所确定等价类的模糊集形式。给定$R$上的一个模糊划分$\theta$,利用上近似$\theta^*$和下近似$\theta_*$表达模糊集合$F$,称$(\theta^*(F),\theta_*(F))$为模糊-粗糙集。则模糊上、下近似[3]定义如下:

$$M_i=\mu_{\theta^*(F)}(F_i)=\sup_x(\min(\mu_{F_i(x)},\mu_{F(x)}))\ \forall i$$

$$m_i=\mu_{\theta_*(F)}(F_i)=\inf_x(\max(1-\mu_{F_i(x)},\mu_{F(x)}))\ \forall i$$

性质1 $\forall x_j\in U$,若$x_j\notin M_i$,则$\exists l\in\{1,2,\cdots,k\}\wedge l\neq j$,有$x_j\in M_l$。

性质2 $\forall x_j\in U$,若$x_j\in m_i$,则$\forall l\in\{1,2,\cdots,k\}\wedge l\neq j$,有$x_j\in M_l$且$x_j\notin M_l$。

定义2(模糊隶属度矩阵) $X=\{x_1,x_2,\cdots,x_n\}$为论域U内待分类对象集,聚类数目为k,则$x_j(j=1,2,\cdots,n)$对聚类中心$v_i(i=1,2,\cdots,k)$的隶属度μ_{ij}为:

$$\mu_{ij}=\begin{cases}1, & x_j\in M_i\wedge x_j\in m_i\\ 1/H, & x_j\in M_i\wedge attach(x_j,H)\\ 0, & x_j\notin M_i\end{cases}$$

式中,$attach(x_j,H)$表示x_j同时属于H个模糊集的模糊上近似。

由此可得,模糊隶属度矩阵$U=[\mu_{ij}]_{k\times n}$:

到稿日期:2011-12-01 返修日期:2012-01-20 本文受重庆市信息产业发展资金项目(200921011)资助。

刘永忠 男,硕士生,主要研究方向为信息安全、入侵检测、软件工程,E-mail:fangjuny@swu.edu.cn;李欣娣 女,硕士生,主要研究方向为软件工程;李 杨 女,硕士生,主要研究方向为信息安全、软件工程;张为群 男,教授,主要研究方向为软件工程、形式语言、信息安全。

$$U=\begin{Bmatrix}U_1\\U_2\\\vdots\\U_j\\\vdots\\U_k\end{Bmatrix}=\begin{bmatrix}\mu_{11}&\cdots&\mu_{1j}&\cdots&\mu_{1n}\\\mu_{21}&\cdots&\mu_{2j}&\cdots&\mu_{2n}\\\vdots&\ddots&\vdots&\ddots&\vdots\\\mu_{i1}&\ddots&\mu_{ij}&\ddots&\mu_{in}\\\vdots&\ddots&\vdots&\ddots&\vdots\\\mu_{k1}&\cdots&\mu_{kj}&\cdots&\mu_{kn}\end{bmatrix},\mu_{ij}\in[0,1]$$

定义 3(攻击特征)[4]

$$S=\left\{\begin{matrix}sip(d),dip(d),sport(d),dport(d),mask(d),\\protocol(d),service(d),log(d),dur(d),sb(d),db(d),\\count(d),sc(d),dhsc(d),dhsspr(d),dhrr(d)\end{matrix}\right\}$$

为数据连接记录的攻击特征。

式中 $sip(d)$ 和 $dip(d)$ 为数据连接的源 IP 与目的 IP 地址，$sport(d)$ 和 $dport(d)$ 代表连接的源端口号和目的端口号，$mask(d)$ 代表数据连接的子网号，$protocol(d)\in\{TCP,UDP\}$ 和 $service(d)\in\{TELNET,FTP,SMTP,DNS,HTTP,SNMP,TFTP\}$ 分别为数据连接的协议及请求的服务类型，$log(d)\in\{0,1\}$ 代表是否成功登录系统，$dur(d)\in N$ 代表连接持续时间，$sb(d)\in N$ 和 $db(d)\in N$ 分别代表从源端传向目的端的字节数和从目的端传向源端的字节数，$count(d)\in Z^+$ 代表连接次数，$sc(d)\in N$ 代表目标端口与当前连接相同的连接次数，$dhsc(d)\in N$ 代表目标主机和端口与当前连接相同的连接次数，$dhsspr(d)\in[0,100]$ 和 $dhrr(d)\in[0,100]$ 分别代表目标主机相同连接的百分比和含有“SYN”错误的连接所占的百分比。

定义 4(相异匹配测度) $d(x_{lj}^s,v_{li}^s)=\begin{cases}0, & x_{lj}^s=v_{li}^s\\1, & x_{lj}^s\neq v_{li}^s\end{cases}$，$(1\leqslant j\leqslant n,1\leqslant i\leqslant k)$ 称为数据对象 x_j 与聚类中心 v_i 关于第 l 个离散特征值的相异匹配测度。

定义 5(攻击类型与聚类结果映射) $f:Attack\rightarrow V$，$Attack=\{i\mid i\in Z\wedge 1\leqslant i\leqslant 6\}$，$V=\{f(i)\mid i\in Z\wedge 1\leqslant i\leqslant 6\}$，$\forall Attack_1,Attack_2\in Attack$，$Attack_1\neq Attack_2\Leftrightarrow f(Attack_1)\neq f(Attack_2)$ $f(1)$=正常数据连接(Normal 类)，$f(2)$=DOS 类攻击，$f(3)$=Probing 类攻击，$f(4)$=R2L 类攻击，$f(5)$=U2R 类攻击，$f(6)$=其它未知类型攻击(Other 类)。

定义 6(检测结果) $O_i=\{N,A\}$，$i\in[1,T]$ 称为第 i 个子分类器的检测结果；T 表示子分类器的总数，$O_i(N)=\{x_j\mid j\in[1,n]\}$ 代表正常数据的集合，$O_i(A)=\{x_j\mid j\in[1,n]\}$ 代表异常数据的集合；且理想检测状态下应满足如下性质：

性质 3 $\forall i,j\in[1,T]$，$O_i(N)=O_j(N)\wedge O_i(A)=O_j(A)$。

性质 4 $\forall x_j\in X$，若 $x_j\in O_i(A)$，则 $x_j\notin O_i(N)$。

定义 7(攻击逻辑值)

$O_i(x_j)=\begin{cases}1 & x_j\in O_i(A),\forall x_j\in X\\0 & x_j\in O_i(N),\forall x_j\in X\end{cases}$ 称为数据记录 x_j 的攻击逻辑值。

定义 8(加权集成) 设 $W=\{w_i\mid i\in[1,T]\}$ 为 T 个子分类器分类结果所占的权重，$0<w_i<1$ 且 $\sum_{i=1}^{T}w_i=1$，则：

$\forall x_j\in X$，若 $\sum_{i=1}^{T}w_i*O_i(x_j)\geqslant 0.5$，则 $x_j\in O(A)$；

$\forall x_j\in X$，若 $\sum_{i=1}^{T}w_i*O_i(x_j)<0.5$，则 $x_j\in O(N)$；

2.2 线性组合距离

设 $X=\{x_1,x_2,\cdots,x_n\}$ 为待聚类分析的数据连接记录全体，$V=\{v_1,v_2,\cdots,v_k\}$ 为聚类中心，则 x_j 与 v_i 之间的线性组合距离为：

$$D(x_j,v_i)=\sum_{l=1}^{t}\omega_l d(x_{jl}^s,v_{il}^s)+\sum_{l=t+1}^{S}\omega_l||x_{jl}^c-v_{il}^c||^2 \quad (1)$$

式中，$x_j=[x_j^s,x_j^c]^T$ 表示 x_j 的 S 个特征值，其中包括离散型特征 $x_j^s=[x_{j1}^s,\cdots,x_{jt}^s]$ 及连续型特征 $x_j^c=[x_{j,t+1}^c,\cdots,x_{jS}^c]$，$v_i=[v_{i1}^s,\cdots,v_{it}^s,v_{i,t+1}^c,\cdots,v_{iS}^c]^T$ 为聚类中心；指标权重 ω_l 由 ReliefF 技术[5]迭代确定，且其满足 $0<\omega_l<1$ 及 $\sum_{l=1}^{S}\omega_l=1$。由此可得出基于指标间线性组合距离的 FRS-FCM 算法的聚类准则函数[6]为：

$$J(U,V)=\sum_{j=1}^{n}\sum_{i=1}^{k}(u_{ij})^m D^2(x_j,v_i) \quad (2)$$

式中，m 是模糊度指标，根据经验通常取值为 2。然后运用 Lagrange 乘子法，得到改进 FCM 算法的隶属度计算公式为：

$$\mu_{ij}=\frac{1}{\sum_{z=1}^{k}\left[\frac{D(x_j,v_i)}{D(x_j,v_z)}\right]^{1/(m-1)}},i=1,\cdots,k,j=1,\cdots,n \quad (3)$$

聚类中心计算公式为：

$$v_i=\frac{\sum_{j=1}^{n}(\mu_{ij})^m x^j}{\sum_{j=1}^{n}(\mu_{ij})^m},i=1,\cdots,k \quad (4)$$

3 基于 FRS-FCM 的集成入侵检测方法

3.1 FRS-FCM 算法

传统 FCM 算法对初值的依赖性过大且欧氏距离对噪声数据比较敏感，只适用于处理数值型数据及特征空间为超球结构的数据集，对超立方体结构、超椭球结构的数据集效果不太理想。根据聚类应使类内距离尽量小、类间距离尽可能大的原则，本文提出使用基于特征加权的线性组合距离来评价数据记录之间的相似或相异程度。在确定改进 FCM 算法的聚类数目 k、模糊隶属度矩阵 U、模糊度参数 m 和容许误差 ξ 后，聚类的过程就是求取相邻两次聚类准则函数 $J(U,V)$ 值之差小于容许误差 ξ 的过程；算法具体流程如图 1 所示。

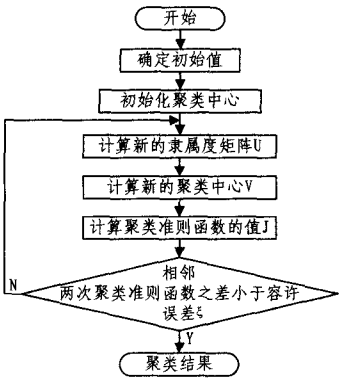

图 1 FRS-FCM 算法聚类流程

3.2 基于 FRS-FCM 的集成入侵检测方法

本文针对当前高速网络环境、海量数据的情况，提出了如图 2 所示的基于 FRS-FCM 算法的集成入侵检测[7]方法模型，该模型具体检测算法形式化描述如下：

Step1 准备(Prepare) getDataSet() 表示数据集抽取，preProcess(d) 表示对数据记录 d 进行预处理，RISC(D) 表示

数据约简，则有：

Input：网络数据

D=getDataSet()

For each d in D

preProcess(d)

End For

X=RISC(D)

Output：$X=\{x_1, x_2, \cdots, x_n\}$

Step2　求取模糊上下近似集合(getApproximateSet)：

Input：X，F

$M=\phi$，$m=\phi$

For each F_j in F

$m_i=getInf(F_j)$

$M_i=getSup(F_j)$

$M=M\cup M_i$

$m=m\cup m_i$

End For

Output：M，m

Step3　求取模糊隶属度矩阵(getFuzzyMatrix)：

Input：M，m，X，V

For each x_t in X

For each v_j in V

If($x_t \notin M_{ji}$) $\mu_{jt}=0$

Else If($x_t \in m_{ji}$) $\mu_{jt}=1$

Else If($x_t \in M_{ji} \wedge Attach(H)$) $\mu_{jt}=1/H$

End If

End For

End For

Output：$U=[\mu_{jt}]_{k\times n}$

Step4　聚类(getCluster)：getInitValue()表示得到聚类所需初始值，getNewMatrix()表示求取新的模糊隶属度矩阵，getNewCenter()表示求取新的聚类中心，getFunValue()表示求取聚类准则函数的值。

Input：U，X

getInitValue()

While($|J-J'|\geqslant\xi$)

$J'=J$

U=getNewAtrix()

V=getNewCenter()

J=getFunValue()

End while

Output：$V=\{v_1, v_2, \cdots, x_k\}$

Step5　检测(Detect)：

Input：X，V，W

$O_{GABP}(N,A)=GABP\_Detect(V)$

$O_{SCG\_BP}(N,A)=SCGBP\_Detect(V)$

$O_{SVM}(N,A)=SVM\_Detect(V)$

For each x in X

If($W_{GABP} * O_{GABP}(x)+W_{SCG\_BP} * O_{SCG\_BP}(x)+W_{SVM} * O_{SVM}(x)\geqslant 0.5$)

$x\in O(A)$

Else $x\in O(N)$

End If

End For

Output：O(N，A)

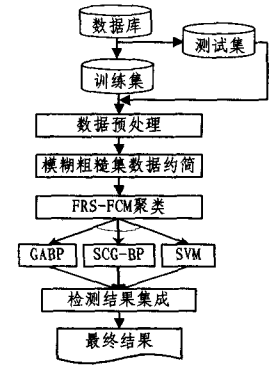

图 2　基于 FRS-FCM 算法的集成入侵检测模型

4　实验

本文采用来自于 DARPA 检测评估计划的 KDD Cup 1999 数据集来验证基于 FRS-FCM 算法的集成入侵检测模型的可行性与有效性。该数据集分为训练数据集与测试数据集两大部分，包含了大量的网络入侵数据和正常流量数据，其中训练数据集包含 23 种入侵行为，且训练数据集中每条数据连接均已标识为正常或特定的攻击类型；而测试数据集包含 38 种入侵行为，每条数据连接攻击类型未知。部分攻击行为只包含在测试数据集而没有包含在训练数据集中，有利于检测算法通过对训练数据集的学习，识别新型攻击行为的能力。训练集与测试集中每条数据连接记录都包含 41 个属性，其中离散属性 7 个，连续属性 34 个。

在本文的仿真实验中，首先从 KDD Cup 10%数据集的训练数据集 kddcup. data_10_percent. gz 中随机抽取 21000 条数据来建立一个训练集 TR，但考虑到 PRB、R2L 和 U2R 在整个数据集中所占比例较小，把 kddcup. data_10_percent. gz 中所有的 PRB、R2L 和 U2R 攻击都选到本训练数据集 TR 中，以保证对整个入侵检测模型训练得更全面；而且真实网络环境中正常连接的比例远大于异常连接，所以最后确定的训练数据集[10] TR 如表 1 所列。

表 1　训练数据集(TR)结构

连接类型	连接数目	所占百分比
Normal	10000	42.62%
Dos	5715	27.21%
PRB	4107	19.56%
U2L	1126	5.36%
U2R	52	0.25%

在使用训练集 TR 对整个入侵检测模型进行的训练结束后，在 KDD Cup 10%数据集的测试数据集 kddcup. newtestdata. unlabeled_10_percent. gz 中按正常与异常比 19:1 随机抽取 5 个测试数据集{TS1，TS2，TS3，TS4，TS5}，使得每个测试数据集中都包含 2000 条连接，然后采用这 5 个测试数据集对本文提出的基于 FRS-FCM 算法的集成入侵检测模型与基于传统 FCM 算法的检测模型、基于 GABP 的检测模型进行对比实验，实验主要从检测率、误报率与对低频攻击的检测率(简称低检率) 3 个方面进行对比，实验结果如表 2 所列。

通过表 2 对比实验结果可以发现，所提基于 FRS-FCM 算法的集成入侵检测方法提高了检测率，降低了误检率，并对低频攻击的检测效果有一定的改善，从而验证了本算法是可行的、有效的。

表 2　基于 FRS-FCM 的集成入侵检测模型与传统 FCM、GABP 检测模型的检测率、误报率与低频攻击检测率对比

实验结果	连接总数	正常连接数	异常连接数	FRS-FCM			FCM			GABP		
				检测率(%)	误检率(%)	低检率(%)	检测率(%)	误检率(%)	低检率(%)	检测率(%)	误检率(%)	低检率(%)
TS1	2000	1900	100	85	5.1	89.7	82	5.3	79.9	83	7.8	65.3
TS2	2000	1900	100	90	4.8	90.1	85	7.1	81.3	90	8.2	72.6
TS3	2000	1900	100	86	6.7	79.5	88	8.8	80.2	76	8.8	78.3
TS4	2000	1900	100	85	5.4	83.9	83	9.1	75.8	83	6.7	69.4
TS5	2000	1900	100	86	7.4	86.8	85	7.6	78.4	81	8.1	70.1

结束语　本文针对当前入侵检测方法存在的不足，提出基于模糊粗糙集与 ReliefF 技术的 FRS-FCM 算法，并将此算法运用到集成入侵检测中。最后，利用 KDD Cup 1999 数据集进行实验，验证了基于 FRS-FCM 算法的集成入侵检测方法能够在降低误检率的同时提高检测率与泛化能力，并对低频攻击的检测率有较大的提高。但本方法在确定聚类中心个数以及子分类器检测结果的集成权重时还存在主观因素，而且对检测报警信息并没有进行相关处理，因此下一步的研究工作主要包括以下两个方面：(1)采用机器学习算法确定初始聚类中心的个数以及子分类器检测结果的集成权重，并对集成权重进行动态学习更新；(2)对报警信息进行相关性和聚类处理，增强报警针对性，以减少网络管理员的工作负担。

参 考 文 献

[1] Tsai C-F, Hsu Y-F, Lin C-Y. Intrusion detection by machine learning: A review [J]. Expert Systems with Applications, 2009, 36: 11994-12000

[2] Wang Gang, Hao Jin-xing, Ma Jian, et al. A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering [J]. Expert System with Applications, 2010, 37: 6525-6232

[3] 袁妍，洪晓光. 基于模糊-粗糙集的移动对象 k 近邻预测[J]. 计算机科学，2008，35(2)：140-143

[4] 赵越，张为群. 一种基于 CFCM 的集群入侵检测方法的研究[J]. 计算机科学，2010，37(6)：176-178

[5] 李洁，高新波. 基于特征加权的模糊聚类新算法[J]. 电子学报，2006，34(1)：89-95

[6] 王骏，王士同. 基于混合距离学习的双指数模糊 C 均值算法[J]. 软件学报，2010，21(8)：1878-1888

[7] 徐冲，王汝传. 基于集成学习的入侵检测方法[J]. 计算机科学，2010，37(7)：217-219

[8] 吴春琼. 基于神经网络与遗传算法的入侵检测研究[J]. 计算机安全，2010，11：25-27

[9] 张义荣. 一种基于粗糙集属性约简的支持向量异常入侵检测方法[J]. 计算机科学，2006，33(6)：64-71

[10] 杨德刚. 基于模糊 C 均值聚类的网络入侵检测算法[J]. 计算机科学，2005，32(1)：86-91

[11] 肖敏，柴蓉，杨富平，等. 基于可拓集的入侵检测模型[J]. 重庆邮电大学学报：自然科学版，2010，22(3)：345-349

(上接第 100 页)

管理，为每组设定组管理员。由组管理员选择成员节点，由发送者编排路径节点通信顺序。本文证明了无环分组路由选择机制的匿名度与系统中非泄密节点成员呈比例的关系，并分析了相同路径长度下的下一跳路由选择机制的匿名度，将两者进行了对比和测试分析。实验数据表明，在通信路径长度相同的情况下，无环分组的路由选择策略比下一跳路由选择策略的匿名性能要好，并且在路径长度 $n>10$ 时，下一跳路由选择策略的匿名性能不会再提高，而无环分组的路由选择机制的匿名性能却一直增加。直到 $n>30$ 时，匿名性能才趋于定值，进一步完善了 P2P 网络的安全性。

参 考 文 献

[1] 欧中洪，宋美娜，战晓苏，等. 移动对等网络关键技术[J]. 软件学报，2008，19(2)：404-418

[2] Chaum D. Untraceable electronic mail, return addresses, and digital pseudonyms [J]. Communications of the ACM, 1981, 4(2): 84-88

[3] Jakobsson M. A practical mix[C]// EUROCRYPT'98. 1998: 448-461

[4] 高蕾，李大兴. 基于 P2P 网络的匿名通信系统[D]. 青岛：山东大学，2008

[5] 孙黎，王小刚. 基于结构化 P2P 的可控匿名通信系统的研究[J]. 科学技术与工程，2010，5(1)：306-310

[6] 陆天波，时金桥，程学旗. 基于互联网的匿名技术研究[J]. 计算机科学与探索，2009，3(1)：35-42

[7] M'Raihi D, Pointcheval D. Distributed Trustees and Revocability: A Framework for Internet Payment[C]// Lecture Notes in Computer Science. 1998, 1645: 28-50

[8] Zhang F, Zhang F T, Wang Y. Fair electronic cash systems with multiple banks based on ACJT group blind signature[J]. Journal of Wuhan University of Technology, 2000, 32(5): 849-852

[9] 邓琳，谢鲲，李仁发. P2P 匿名通信系统的匿名度量及协议研究[D]. 长沙：湖南大学，2009

[10] Serjantov A, Danezis G. Towards an information theoretic metric for anonymity[C]// Proceedings of Privacy Enhancing Technologies Workshop. 2003: 41-53

[11] Sander T, Ta-Shma A. Flow control: A new approach for anonymity control in electronic cash systems [J]. Conference on Computational Intelligence and Security, 1999, 1(1): 354-379P

[12] 江丽，徐红云. 基于组群的匿名通信协议研究与探讨[J]. 计算机工程与应用，2008，44(9)：125-128

[13] Camenisch J, Maurer U, Stadler M. Digital payment systems with passive anonymity-revoking trustees[J]. Journal of Computer Security, 1997, 5(1): 69-89