一种基于CFCM的集群入侵检测方法的研究

赵 越 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 将网络数据流聚类来实现负载平衡已经被广泛应用于集群入侵检测方法中。将相关性思想引入传统模糊C-均值聚类算法(FCM),给出数据流逻辑距离公式,提出了一种相关模糊C-均值聚类算法(CFCM)。最后,将此算法应用于集群入侵检测方法中,利用KDD Cup 1999数据集进行实验,验证其可行性及准确性。

关键词 相关性思想,CFCM,集群入侵检测方法
中图法分类号 TP309.5 **文献标识码** A

Research Based on a Method of CFCM Clustering Intrusion Detection

ZHOU Yue ZHANG Wei-qun
(College of Computer and Information Science, Southwest University, Chongqing 400715, China)
(Intelligent Software and Software Engineering Research Institution, Chongqing 400715, China)

Abstract Clustering Web flows to attain Load Balancing is widely used in Clustering Intrusion Detection approach. Correlated ideas were imported in the traditional fuzzy C-means clustering algorithm (FCM), defined logical distance formula, and presented a Correlated fuzzy C-means clustering algorithm(CFCM). At last we applied this approach in Clustering Intrusion Detection approach, and used KDD Cup 1999 data set to do experiment. The results demonstrated the viability and effectiveness of our approach.

Keywords Correlated ideas, CFCM, Clustering Intrusion Detection approach

入侵检测技术作为防火墙等传统手段的必要补充,可以在不影响性能的情况下对网络进行检测。一个完备的入侵检测系统是基于主机的入侵检测系统与基于网络的入侵检测系统的结合。但传统的单个IDS也存在缺点,即随着网络带宽的增加易产生数据超载现象,从而引起掉包,极大地影响了检测的准确率。

传统的模糊C-均值聚类算法[1],考虑了每一个待识别的对象具有亦次亦彼的性质,但这种柔性的聚类划分方法没有考虑到数据流之间的相关性,从而影响聚类划分的准确率。

本文综合分析以上两方面的优势与不足,将数据流相关性的思想引入传统模糊C-均值算法(FCM)中,提出一种相关模糊C-均值聚类算法(CFCM),将此算法应用于集群入侵检测方法中,既避免数据超载现象,又保证了检测的准确率。

1 基本概念

1.1 数据流相关性

本文首先采用TCPDUMP,将网络中传送数据流的"头"完全截获下来分析处理,处理结果用网络连接记录表示。首先根据文献[2]判断协议类型属于TCP还是UDP。分析协议类型后,将网络数据流连接记录的格式归结如表1所列。

表1 数据流连接记录的格式

内容 / 占有bit数	源IP地址	目的IP地址	源端口号	目的端口号	连接持续时间(秒)	服务类型
bit数(b)	32	32	16	16	9	8

由表1可见,一个经处理得到的连接记录中包含了描述该数据流的基本属性和辅助信息。基本属性如源IP地址、目的IP地址、目的端口号等,辅助信息如连接持续时间等。

本文首先将连接记录中的基本属性抽象成一个集合。设任意数据流 d, $F=\{sip(d), dip(d), mask(d), dport(d), protocol(d), service(d)\}$ 为其连接记录基本属性集合。其中 $sip(d)$ 和 $dip(d)$ 代表数据流的源IP地址与目的IP地址, $dport(d)$ 代表目的端口号。$mask(d)$ 代表数据流子网号, $protocol(d)$ 和 $service(d)$ 分别为数据流的协议及请求的服务。f 是 F 中的任意一个属性,任意待聚类数据流 $b_j, j\in\{1,N\}$,已存在数据流 $a_i, i\in\{1,M\}$, def 表示以 $d_{port}(a_i)$ 为变量的函数,数据流 b_j 与 a_i 的匹配有如下定义。

定义1(匹配) $\forall f\in F$,判断 $b_{jf}=a_{if}$,对于 $dport(a_i)$ 与 $dport(b_j)$ 还需判断 $def(d_{port}(a_i))=d_{port}(b_j)$ 和它们是否来自同一管理者,以上操作称为数据流 b_j 与 a_i 关于属性 f 的匹配操作。

设 $P_f(b_j,a_i)$ 代表数据流关于属性 f 匹配的结果函数,是

到稿日期:2009-07-21 返修日期:2009-09-30 本文受重庆市自然科学基金重点项目"软件测试技术和方法研究"(CSTC,2006BA2003)资助。
赵 越(1985—),女,硕士生,主要研究方向为软件工程,E-mail:zy850302@swu.edu.cn;张为群(1950—),男,教授,硕士生导师,CCF会员,主要研究方向为软件工程、软件测试与可信计算。

Bool 型值。Cor_{ijf} 代表返回函数值。数据流 b_j 与 a_i 匹配具有如下性质：

性质 1 $P_f(b_j,a_i)=\begin{cases}Ture, b_{jf}=a_{if}\\ False, b_{jf}\neq a_{if}\end{cases}$。

性质 2 $Cor_{ijf}=\begin{cases}1, P_f(b_j,a_i)=\text{Ture}\\ 0, P_f(b_j,a_i)=\text{False}\end{cases}$，

$Cor_{ijF}=\sum_{\forall f\in F}Cor_{ijf}$。

数据流 b_j 与 a_i 的在属性集 F 上进行匹配时原则如下：

(1)IP 地址

若 $f'\subset F, f'=\{s_{ip}(),d_{ip}(),mask()\}$，

$Cor_{ijf'}=\begin{cases}1, P_{f'}(b_j,a_i)=\text{Ture}\\ 0, P_{f'}(b_j,a_i)=\text{False}\end{cases}$。

(2)$dport$

$Cor_{ijdport}=\begin{cases}1, P_{dport}(b_j,a_i)=\text{Ture}\\ 0, P_{dport}(b_j,a_i)=\text{False}\end{cases}$；

若 $def(d_{port}(a_i))=d_{port}(b_j), Cor_{ijport_{def}}=1$；

当 $b_{jdport}\neq a_{idport}$，设有管理者 $A_k, k\in[1,\infty), set_k(dport())\subset A_k$，若 $(d_{port}(a_i)\cup d_{port}(b_j))\subset s_k(d_{port}())$，那么 $Cor_{ijport_{fuc}}=1$。

(3)protocol

$Cor_{ijprotocol}=\begin{cases}1, P_{protocol}(b_j,a_i)=\text{Ture}^{[4]}\\ 0, P_{protocol}(b_j,a_i)=\text{False}\end{cases}$。

(4)service

$Cor_{ijservice}=\begin{cases}1, P_{service}(b_j,a_i)=\text{Ture}\\ 0, P_{service}(b_j,a_i)=\text{False}\end{cases}$。

定义 2(数据流相关性) 对于数据流 b_j 与 a_i，若 $\forall f\in F, Cor_{ijF}>0$，称数据流 b_j 与 a_i 具有相关性[3]。

1.2 逻辑距离

定义 3(逻辑距离) $\forall f\in F, d_f(b_{jf},a_{if})$ 表示数据流 b_j 到 a_i 关于属性 f 的逻辑距离，描述其进行相关性匹配时的差异程度。

定理 1 对于 $\forall f\in F$，有 $d_f(b_{jf},a_{if})+Cor_{ijf}=1$。

设 $d_f(b_{jf},a_{if})$ 表示数据流 b_j 和 a_i 在属性集 F 上的逻辑距离，进行匹配时其距离公式可以表示为：

$$d_f(b_{jf},a_{if})=\begin{cases}1, Cor_{ijf}=0\\ 0, Cor_{ijf}=1\end{cases}\tag{1}$$

设 $d_1(b_{jf'},a_{if'})$ 表示数据流 b_j 和 a_i 端口号函数相关性的逻辑距离，$f''=d_{port}(), f''\in F$，逻辑距离公式如下：

$$d_1(b_{jf''},a_{if''})=\begin{cases}1, Cor_{ijport_{def}}=0\\ 0, Cor_{ijport_{def}}=1\end{cases}\tag{2}$$

设 $d_2(b_{jf'},a_{if'})$ 表示数据流 b_j 和 a_i 功能相关的逻辑距离，在判断逻辑距离时，采用如下公式：

$$d_2(b_{jf'},a_{if'})=\begin{cases}1, Cor_{ijport_{fuc}}=0\\ 0, Cor_{ijport_{fuc}}=1\end{cases}\tag{3}$$

当数据流 b_j 和 a_i 相关性为 0 时，其逻辑距离达到最大值。将数据流 b_j 与 a_i 的各个属性值按照 $d_{ip}, d_{port}, d_{protocol}, d_{service}$ 这 4 项指标汇总，设 $d(b_{jip}-a_{iip})$ 表示数据流 b_j 与 a_i 的 IP 地址逻辑距离，公式如下：

$$d(b_{jip}-a_{iip})=\sqrt{\frac{\sum_{\forall f_1\in f}d_{f_1}(b_{jf_1},a_{jf_1})^2}{3}}\tag{4}$$

设 $d(b_{jport}-a_{iport})$ 表示数据流 b_j 与 a_i 的端口号逻辑距离，逻辑距离公式如下：

$$d(b_{jport}-a_{iport})=\sqrt{\frac{d_f(b_{jf'},a_{if'})^2+d_1(b_{jf''},a_{if''})^2+d_2(b_{jf''},a_{if''})^2}{3}}\tag{5}$$

设 $d(b_{jprotocol}-a_{iprotocol})$ 与 $d(b_{jservice}-a_{iservice})$ 分别表示数据流 b_j 与 a_i 的协议逻辑距离及请求服务逻辑距离。当 f 分别取 $protocol()$ 和 $service()$ 时，根据式(1)可以计算出其逻辑距离。即：

当 $f=protocol()$ 时，$d(b_{jprotocol}-a_{iprotocol})=d_f(b_{jf},a_{if})$ (6)

当 $f=service()$ 时，$d(b_{jservice}-a_{iservice})=d_f(b_{jf},a_{if})$ (7)

1.3 改进的欧氏距离

由于攻击类型不同，且在判断数据流与聚类中心的逻辑距离时，各个属性发挥的作用不同，因此给各个属性加权[5]，以此体现出每个属性的特征。

改进的欧氏距离表示了数据流各属性间的距离，设给各属性加权值 μ_k，定义如下：

$$d(b_j-a_i)=\sqrt{\sum_{\forall k\in A}\mu_k(b_{jk}-a_{ik})^2}\tag{8}$$

式中，$A\in\{ip,port,protocol,service\}$

式(8)将数据流 b_j 与数据流 a_i 的 4 个属性的逻辑距离加权求和，其中 $(\mu_{ip}=\mu_{port})>(\mu_{ip}=\mu_{ip})$ 且权值之和为 1。

CFCM 的目标函数为：

$$J_m(U,C)=\sum_{j=1}^{N}\sum_{i=1}^{c}[u_i(x_j)]^m(\sqrt{\sum_{\forall k\in A}\mu_k(x_{jk}-c_{ik})})^2\tag{9}$$

FCM 的目的是求目标函数式(1)最小值，则 J_m 对 c_{ik} 和 J_m 对 $u_i(x_j)$ 求偏导得出隶属度函数及聚类中心函数为：

$$u_i(x_j)=\frac{[1/\|\mu_k(x_{jk}-c_{ik})\|^2]^{1/(m-1)}}{\sum_{j=1}^{N}\sum_{i=1}^{c}[1/\|\mu_k(x_{jk}-c_{ik})\|^2]^{1/(m-1)}}\tag{10}$$

$$c_{ik}=\frac{\sum_{i=1}^{N}\sum_{i=1}^{c}[u_i(x_j)]^m x_j}{\sum_{j=1}^{N}\sum_{i=1}^{c}[u_i(x_j)]^m}\tag{11}$$

聚类中心 $c_i=\sum_{\forall k\in A}c_{ik}/4$ (12)

随着数据流 b_j 的增多能够实时地计算其与聚类中心 a_i 的欧氏距离，将阈值 λ 取改进欧氏距离的最大值与最小值之和的平均值，即：

$$\lambda=\frac{1}{2}[\min(d(b_j-a_i))+\max(d(b_j-a_i))]\tag{13}$$

2 基于 CFCM 的集群入侵检测方法

2.1 CFCM

CFCM 中聚类权重 w，描述原有聚类中数据流数目；聚类数目 c，描述原有聚类数目；th 为预先设定的阈值。设有聚类 A 和 B，聚类中心分别为 a_i 和 b_j。设 C 表示多次迭代后的聚类数目，$C_{\max}$ 为聚类数目上限。

输入：初始化后的 c 个聚类，数据流 b_j。

输出：根据 CFCM 优化的 c' 个聚类。

step1 提取 a_i 和 b_j 的属性集 F。

step2 对 $\forall f\in F$，求其相关性，若 $Cor_F=0$，转 step6。

step3 若 $Cor_F=1$，初始化隶属矩阵 U，并使其满足式 $\sum_{i=1}^{n}u_{ij}=1, \forall j=1,\cdots,n$。根据式(1)－式(8)计算出 $d(b_j-a_i)$。

step4 若 $d(b_j-a_i)<\lambda$，将数据流 b_j 加入 A 中，计算 $u_i(b_j)$，更新隶属度矩阵 U，根据式(11)计算出新的聚类中心，$w=w+1$，否则转 step7。

step5　计算 $J_m(U,C)$，若 $J_m(U,C)<th$，算法终止，返回结果 c' 个聚类，否则转 step1。

step6　新建聚类 B，$w=1$，$c=c+1$。

step7　若 $C>C_{max}$，且 $C-C_{max}=n$，则删除聚类中数据流数目最小的 n 个，转 step1。

算法的时间复杂度为 $O(n)$。

2.2　基于 CFCM 的集群入侵检测方法

本文提出的相关模糊 C-均值聚类算法是建立在集群入侵检测系统模型(见图 1)基础上的。

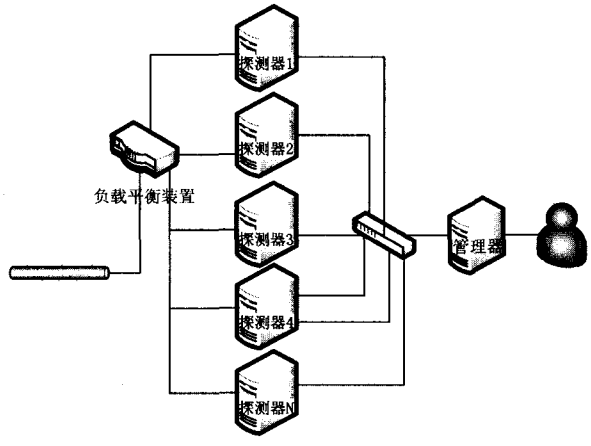

图 1　集群入侵检测系统模型[6]

在负载平衡装置中，将应用聚类算法将数据流实时聚类，并分发到各探测器中，负载平衡装置中基于聚类技术的集群入侵检测过程如图 2 所示。

step1　采集网络数据：提取数据流的属性组 F。

step2　数据实例标准化：$x_{ij}=\frac{x_{ij}-\overline{x_j}}{S_j}$，其中 $i=1,2,\cdots,n,j=1,2,\cdots,m,\overline{x_j}=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$，$s_j=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x_j})^2}$[1]。

step3　采用基于模型的聚类方法进行初始化聚类。

step4　采用 CFCM 对初始化聚类进行优化，通过 $J_m(U,C)<th$ 判断聚类是否合理，若合理，则完成分类，否则转 step3。

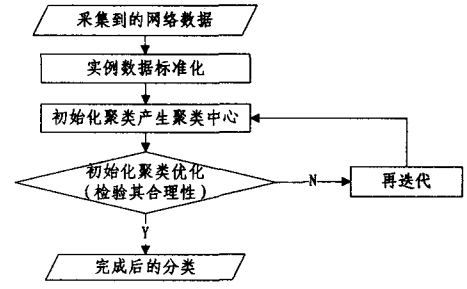

图 2　基于聚类的入侵检测过程

3　实验

本文采用来自于 DARPA 检测评估计划的 KDD Cup 1999 数据集上的数据来验证 CFCM 的有效性。该数据集包括了大部分网络环境下模拟的各种入侵，约 500 万条连接记录的 7 个星期的网络流量，其中每个连接实例包含 41 个属性，且均已标识为正常或特定的攻击类型。

本文采用 MIT Lincoln 实验室的测试数据作为原始数据，采用测试工具 tcpreplay 产生数据流量[7]，测试主机的基本配置为处理器：2.40GHz；内存：512MB；硬盘：80GB；操作系统：Windows2000。负载平衡装置配置为 CPU2.66GHz，内存 512M，网卡为千兆以太网卡；探测器配置同负载平衡装置。表 2 为基于 CFCM 的入侵检测仿真实验结果与基于传统 FCM 的入侵检测结果对比情况。C 代表 CFCM，F 代表传统的 FCM。

表 2 的性能指标反映了 CFCM 对于入侵行为的检测能力优于传统 FCM。在仿真试验中，入侵实例数与正常实例数之比为 1∶10，检测率达到了 67%，误检率为 1.117%，可见此算法对于异常入侵检测是可行的、有效的。

表 3 反映了在探测器数目不同的情况下 CFCM 的集群入侵检测系统模型的检测率。

从表 3 中可以发现，对于集群入侵检测系统，检测率随着探测器数目的增多而增大，可见，集群入侵检测系统提高了入侵检测的性能。

表 2　基于 CFCM 的入侵检测仿真实验结果与基于传统 FCM 的入侵检测结果对比

实验结果	实例总数	正常实例数	入侵实例数	C 正常实例数	C 入侵实例数	F 正常实例数	F 入侵实例数	F 检测率(%)	F 误检率(%)	C 检测率(%)	C 误检率(%)
数据集 1	1100	1000	100	993	53	999	49	49	0.1	53	0.7
数据集 2	1100	1000	100	995	100	1000	41	41	0	100	0.5
数据集 3	1100	1000	100	994	48	986	100	400	1.4	48	0.6
数据集 4	1100	1000	100	997	50	994	50	50	0.6	50	0.3
数据集 5	1100	1000	100	1000	100	1000	100	100	0	100	0
数据集 6	1100	1000	100	954	51	944	20	20	5.6	51	4.6
平均值	1100	1000	100	988.83	67	999.83	60	60	1.283	67	1.117

表 3　不同探测器数目下系统的攻击检测率

网络流量(Mbps)	100	200	300	400	500	600	700	800
1 个探测器攻击检测率(%)	100	90	80	60	50	30	20	7
2 个探测器攻击检测率(%)	100	100	90	70	60	40	25	10
3 个探测器攻击检测率(%)	100	100	100	90	75	60	45	20
4 个探测器攻击检测率(%)	100	100	100	100	90	75	60	40
5 个探测器攻击检测率(%)	100	100	100	100	100	85	60	45
6 个探测器攻击检测率(%)	100	100	100	100	100	100	80	65

结束语　本文针对将聚类算法运用在集群入侵检测方法中未考虑数据流相关性的问题，利用 CFCM 进行入侵检测。最后，利用 KDD Cup1999 数据集对该算法进行实验。通过实验，验证了将 CFCM 应用于集群入侵检测方法中能够提高检测准确率。

参 考 文 献

[1]　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)：86-88

[2]　米文涛. 基于网络连接记录的异常检测分类模型研究[D]. 太原：太原理工大学，2003

(下转第 222 页)

在不同的子图，而关联密切的网页在同一子图。在将网页聚类的基础上，可以根据用户的浏览事务对用户聚类，以便针对不同类型的用户给出偏重有所不同的页面显示。依据的原则是用户事务和网页聚类的相似度。

2.3 空间数据挖掘

空间数据是一类具有多维特征，即时间维、空间维以及众多的属性维的数据，描述了复杂系统的状态、系统的性质、系统的空间分布和系统的发展演化。空间数据挖掘指的是从空间数据库保存的海量空间数据中抽取出空间对象之间的相互关系及反映其演化规律的知识。

超图理论在空间数据挖掘中也具有应用价值。将有向超图同面向对象技术相结合所形成的用于表示模式矢量的结构以及子模式之间关系的图形称为超图模型[12]。超图模型支持类、对象、属性和联系等概念。其中属性包括类的性质和对象的性质；联系包括类与类(对象与对象)之间的连接和关系。超图模型可以表达模式的复杂结构和关系，并将模式同面向对象的相关概念联系在一起，其中包括类、继承、聚合、概括(支持超类)、多层次关系(超类间的联系)以及复合类等。

之所以将超图引入空间数据挖掘，是因为空间数据蕴含有复杂的规律和知识，而超图模型能够用图的方式抽象地描述这种复杂性，从而将空间数据的组织结构予以简化；同时用超图模型还可进行空间数据的可视化表示，便于计算机实现[13]。

超图模型中顶点表示模式矢量的组成(属性、拓扑结构、子模式)，不同的顶点可表示不同抽象层次的模式，顶点之间的有向弧段代表对象类之间的关系。在超图模型中，可以用闭合曲线和赋值的连线表示空间数据的空间维；用点和闭合曲线的赋值以及赋值曲线来表示空间数据的属性维；用属性值的变化以及连线值的变化来表示空间数据的时间维。超图模型还可以表示空间对象之间的层次关系，以及关系的强弱。

空间知识发现的常见的知识类型，如关联规则、聚类和分类规则等，也可利用各种表示逻辑关系的图来表示知识之间的层次和相互依赖。由于超图模型综合了超图和有向图的优点，因此可以将关联规则可视化表示，图中节点表示数据的项，边表示关联关系，规则的可信度和支持度可用不同的颜色和数值来表示。有向图适合于数据和规则数目较少情况下的知识的可视化。

结束语 超图基于图论和集合论，已在数据挖掘研究中得到了运用。超图模型能够利用图的逻辑结构，有效组织和传递数据集的结构、关系和含义，实现关联规则、聚类和分类等知识的获取和表示。超图(特别是有向超图)是 Maradbcm 这一全新关联规则挖掘算法的重要理论基础之一。随着研究的深入，超图理论还将在数据挖掘各研究领域中起到更大的作用。

参 考 文 献

[1] Berge C. Graph and Hypergraph[M]. Amstezdam: North-Holland, 1973

[2] 黄汝激. 超网络的有向超树分析法[J]. 电子科学学刊, 1987, 19(3): 244-255

[3] Agrawal R, Imielinski T, Swami A. Mining Associations between Sets of Items in Massive Databases[C]// Buneman P, Jajodia S, ed. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data(SIGMOD-93). 1993: 207-216

[4] 杨炳儒, 王建新. KDD 中双库协同机制的研究(Ⅰ)[J]. 中国工程科学, 2002, 5(4): 41-51

[5] 杨炳儒, 王建新, 孙海洪. KDD 中双库协同机制的研究(Ⅱ)[J]. 中国工程科学, 2002, 4(5): 34-43

[6] 杨炳儒, 孙海洪. 基于双库协同机制的挖掘关联规则算法 Maradbcm[J]. 计算机研究与发展, 2002, 39(11): 1447-1455

[7] 张蓉. 数据聚类技术的研究[J]. 计算机工程与应用, 2002, 16: 145-147

[8] Hung Shao-shin, Liu Shing-min. Using hypergraph-based clustering scheme for traversal prediction in virtual environments[C]// 2007 First IEEE Symposium on Computational Intelligence and Data Mining. Honolulu, HI, USA, 2007: 429-436

[9] Zheng Yu, Qian Rong. A Hypergraph model for clustering scale-free network[C]// The 27th Chinese control conference. CCC, 2008: 561-565

[10] Liu Yang, Zhang Ya-jie. A new data clustering method for farmland evaluation based on fuzzy-hypergraph model[J]. Journal of Wuhan University of Technology(Information & Management Engineering), 2007, 29(11): 126-128

[11] 李尊朝. 基于网页超图分割的 WEB 聚类法[J]. 纺织高校基础科学学报, 2003, 16(3): 261-264

[12] 杨炳儒, 张德政. 超图模型：基于超图的设计模式描述和复用实现[J]. 计算机工程与应用, 2001, 13: 46-48

[13] 孙连英, 彭苏萍, 张德政. 基于超图模型的空间数据挖掘[J]. 计算机工程与应用, 2002, 11: 30-32

(上接第 178 页)

[3] 赵越. 一种基于相关性的网络数据流聚类算法[J]. 西南师范大学学报：自然科学版, 2009(12 增刊)

[4] Le A, Al-Shaer E, Boutaba R. On Optimizing Load Balancing of Intrusion Detection and Prevention Systems[C]// INFOCOM Workshops 2008. IEEE, 2008, 4: 1-6

[5] 鲁瑞华, 张为群. 模糊加权中值滤波器[J]. 计算机科学, 2006, 6: 186-188

[6] 孙钦东, 张德运, 高鹏, 等. 并行入侵检测系统的负载均衡算法[J]. 小型微型计算机系统, 2004, 25: 2215-2217

[7] 伍海波, 陶滔, 唐启涛, 等. 基于负载均衡的并行入侵检测系统设计[J]. 微计算机信息, 2009, 25: 88-90

[8] 代伟, 刘敏, 余永武. 基于 Ad Hoc 网络的混合入侵检测算法[J]. 重庆工学院学报：自然科学版, 2008, 22(3): 75-78