基于多维聚类挖掘的异常检测方法研究

陈 平[1],宋玉蓉[2],蒋国平[2]
(1.南京邮电大学 计算机学院,江苏 南京 210003;
2.南京邮电大学 自动化学院,江苏 南京 210003)

摘 要:网络异常检测是网络管理中非常重要的课题,因此已在近年来得到广泛研究。人们在该领域提出了许多先进的网络流量异常检测方法,但是自动准确地对网络流量进行分类和识别来发现网络中的异常流量仍然是一个非常具有挑战性的问题。文中提出了一种基于多维聚类挖掘的异常检测方法,通过两个阶段来实现异常检测。第一阶段先通过多维聚类挖掘算法,自动对网络中的流量进行多维聚类,第二阶段通过计算多维聚类的异常度来实现异常检测。通过文中的方法,网络中的异常流量被自动归类到不同的有意义的聚类中,通过对这些聚类进行分析可以发现网络中的异常行为。最后通过实验对算法进行了验证,结果表明该方法能够有效检测网络中的异常流量。

关键词:聚类;异常检测;网络安全
中图分类号:TP309 **文献标识码**:A **文章编号**:1673-629X(2012)07-0136-04

Multidimensional Clustering Based Anomaly Detection Research

CHEN Ping[1], SONG Yu-rong[2], JIANG Guo-ping[2]
(1. College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2. College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Network anomaly detection which is a very important issue in network management has been extensively studied in recent years. Although people in the field made a number of advanced works, the accuracy of automatic classification of network traffic to detect and identify abnormal network traffic is still a very challenging problem. It presents a multidimensional clustering based anomaly detection method, by two stages to achieve anomaly detection. The first phase, through multidimensional clustering algorithms, network traffic is automatically mined into different multidimensional clusters. The second phase calculates the degree of multidimensional clusters to achieve anomaly detection. By this method, the abnormal network traffic is automatically classified into different meaningful clusters, and then these clusters can be used to find network anomalies. Finally, this algorithm was validated through experiments, the results show that the method can effectively identify abnormal network traffic.

Key words: clustering; anomaly detection; network security

0 引 言

随着因特网的不断发展,DDoS 攻击[1],网络蠕虫[2]等在当前网络流量中所占据的比例越来越大。并且每一年都有新的攻击形式,新的蠕虫病毒不断出现,给网络管理带来了严重的挑战。因为 DDoS 攻击、网络蠕虫等发生时,同时会带来流量的异常,可以通过挖掘异常流量并进行分析来发现网络中的异常行为。

人们针对网络异常检测技术已经做了大量研究工作,提出了许多不同方法,如:基于机器学习、基于统计方法学、基于特征/行为的研究方法等。文献[3]提出了一种基于半监督学习的行为建模与异常检测方法,能够自动地选择正常行为模式的种类和样本以建立正常行为模型,检测过程具有较高的可靠性。文献[4]提出了一种结合小波检测和可信度评估的异常检测方法,提高了对突发流量的检测效率。传统异常检测方法通常先采集一些相关数据包进行分析,然后建立合适的网络流量模型,再对它进行识别,因此检测结果会受到历史数据的影响,不能够真实反映当前网络流量的行为特征。

聚类是入侵检测领域中一种无监督异常检测方法[5-8],它可以对网络流数据进行挖掘,从而发现当前网络使用情况的重要特征。聚类方法不需要先验知识,能够发现未知攻击,并且聚类的特性适合处理高

收稿日期:2011-11-29;**修回日期**:2012-03-03
基金项目:江苏省自然科学基金项目(BK2010526);教育部博士点基金项目(20103223110003);南京邮电大学引进人才项目(NY209021)
作者简介:陈 平(1982-),男,硕士研究生,研究方向为信息安全、网络安全;宋玉蓉,教授,研究方向为信息安全、复杂网络、病毒传播;蒋国平,教授,研究方向为复杂动态网络。

速网络中的海量数据,因此近年来得到了广泛的研究。文献[9]提出了基于 k-means 聚类和 ID3 决策树学习算法的无监督异常检测算法,分别从 k-means 聚类算法和 ID3 决策树算法中提取异常分数,结合起来生成最终的异常评分值进行异常检测。文献[10]把用户的连续活动看作数据流,并用数据流上的聚类算法对用户的正常行为进行建模来实现异常检测。文献[11]结合信息熵理论,以整体相似度的聚类质量为聚类合并的策略,提出了一种基于划分和凝聚层次聚类的无监督的异常检测算法。文献[12]提出了一种基于改进的 CURE 聚类算法的无监督异常检测方法,并在建模过程中,提出了一种新的基于超矩形的正常行为建模算法,可以迅速、准确地检测出入侵行为。

上述聚类算法虽然能够发现网络中的异常情况,但是由于网络行为的多样性,如何确定聚类标准,往往是非常困难的。文中通过多维聚类挖掘的方法,以自然层次对网络流数据进行聚类,聚类结果较直观,实验结果表明该方法能有效检测网络中的异常流量。

1　数据流的单维流量聚类

使用五元组(sip, dip, sp, dp, protocol)来定义聚类,即源 IP 地址,目的 IP 地址,源端口号,目的端口号,协议号。这几个字段可以是某个特定的值,也可以是所有可能的值,用(*)来表示。单维规则中只有一个字段的值是确定的,比如单维规则(*,*,*,80,*),就表示所有目的端口为 80 的流量。多维规则也使用五元组来表示,但是多维规则中必须有多于一个字段的值为确定的,比如多维规则(192.168.1.101,*,*,*,80,*),就表示源地址为 192.168.1.101 的用户发出的所有目的端口为 80 的流量。满足一定规则的流量的集合称之为类。当类中流量和总流量比例超过设定的门限值 H 的话,该类就称为显著类。不同的维聚类方法也有所不同:

· IP 维。

采用层次方式来表示,根节点用 * 代表该 IP 地址所有的流量,中间用 8 ~ 32 位的前缀来表示,最下面一层的前缀为 32 位,代表一个具体的 IP 地址。

· 端口维。

采用三层表示,根节点用(*)来表示所有可能的值,第二层使用 high,代表端口号大于 1023 的集合,或者 low 代表小于 1024 的端口号集合,第三层表示具体端口号。

· 协议维。

采用两层表示,根节点用(*)来表示所有可能的值,第二层分为 TCP,UDP 协议。

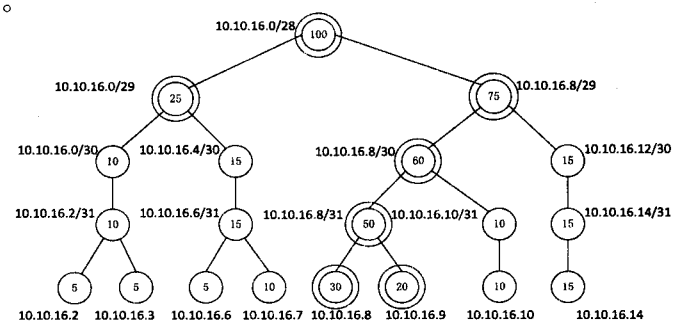

图 1　IP 聚类树

协议维和端口维层次结构较简单,而 IP 维层次结构较复杂。图 1 为 IP 聚类树,图中圈出的节点为显著节点,该 IP 树中包含了大量冗余信息。设流量门限值为 H,压缩门限值为 C,$H = C = 20$,则节点 10.10.16.8/31 的流量为 50,大于 20,为显著节点,但是它的流量完全等于它的两个孩子节点的流量之和,所以该节点为冗余节点。节点 10.10.16.8/30 的流量为 60,大于 20,为显著节点,但是它的流量与它的显著孩子节点流量的差值为 10,小于压缩门限值 20,所以也是冗余节点。

采用以下的压缩算法去除冗余信息,如果两个孩子节点流量都是显著类,则父亲节点肯定是显著类,父亲节点被压缩掉。如果两个孩子节点流量都不是显著类,则只有当父亲节点流量大于门限 H 时,父亲节点流量才为显著类,否则压缩该父亲节点。如果两个孩子节点的流量中有一个是显著类,另一个不是,则只有父亲节点与显著孩子节点流量之差大于门限 C 时,父亲节点为显著类,否则压缩该父亲节点。

设 $C = H = 20$,对图 1 的树结构进行压缩后,可以得到图 2 的树结构。可以看到显著类的数目从 7 个降到了 4 个。

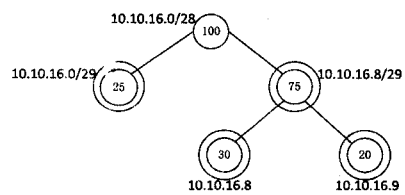

图 2　压缩的 IP 聚类树

2　数据流的多维流量聚类

多维流量聚类树结构,是一种类似于网状的结构,它由多个单维聚类树组合而成。如图 3 所示,右上部分是一个前缀树结构,代表了某个学校(如 E 和 M)的不同部门的流量情况,而左上部分是一个协议的树结构,代表了不同的 TCP 和 UDP 流量。通过组合这两个单维聚类树,就得到了图 3 的下半部分,一个多维聚类树。通过对每一个单维聚类树从上到下、从左到右的

遍历,将一个单维聚类树节点与另一个树的节点结合,就得到了多维聚类树的一个新的节点。例如,将前缀树中的 E 和协议树中的 T 和 U 相结合,就得到了 TE 和 UE,分别代表了学院 E 的 TCP 和 UDP 流量。进一步,通过组合 UE 和 C,可以得到 UC,表示所有系 C 的 UDP 流量。

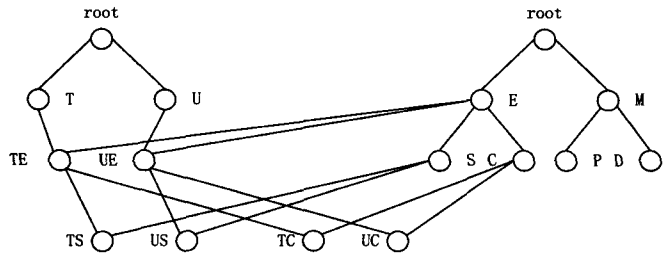

图 3 多维聚类树

可以通过穷举法来得到所有流量超过门限值的多维聚类,即对于每一个多维聚类,就扫描一遍所有的 n 条记录,然后将满足条件的加起来,若流量超过门限值 H,则就得到了一个多维显著聚类。但是这种算法效率太低,在实际中不可行,需要采用一些措施进行优化。

· 在构造多维聚类的时候,总是从父亲节点开始,而若父亲节点不显著,则孩子节点必不显著,因此,只需考虑所有单维父亲节点都是显著类的多维聚类。

· 因为孩子节点派生自父亲节点,父亲节点不匹配的流记录,孩子节点肯定也不匹配。扫描孩子节点时只需在匹配父亲节点的流记录中进行,这样就可以大大减少要扫描的记录。

· 因为端口维和协议维较简单,所以可以先构造端口维和协议维聚类树,再跟 IP 维聚类树组合来提高效率。

3 数据流多维聚类的异常检测方法

经过第一阶段的聚类,按照网络数据流的属性特征对数据流进行聚类,挖掘出了具有特定流量模式的多维聚类。接下来要对得到的聚类进行分析。在检测异常行为方面有许多准则可用,在文献[8]中提出了一种基于流量的判断准则,它将聚类的异常度定义为聚类的实际流量值与聚类的每个单独字段上的流量值乘积的比值。对于聚类 k,它的异常度计算公式为:

$$\text{unexpectedness}(k) = 100 * \frac{p(k)}{\prod_{i=1}^{5} P_i(k)}$$

其中 $p(k)$ 为聚类 k 所占全部流量的百分比,$P_i(k)$ 为聚类 k 中第 i 个字段所占全部流量的百分比。该方法能较好地判断聚类 k 的流量是否异常,但要可靠地判断各种网络攻击行为,如 DDoS,网络蠕虫攻击等,准确率不是很高。

通过对网络中发生 DDoS,网络蠕虫攻击时的异常流量进行分析,可以发现一般网络中发生攻击时,都会带来突发的异常流量,在一段连续的时间内产生大量的数据包,且数据包长度是固定的。因此,在总体表现上,异常流量形成的攻击聚类有下列特点:

· 攻击聚类的密度较大。

· 攻击聚类中的样本数目较大。

综合上述特点,定义聚类的距离标准差 msd 来表示聚类的密度,msd 值越小则聚类的密度越大。N 表示攻击聚类中样本的数目,则聚类的异常度为:

$$msdn(k) = \frac{msd}{N}$$

聚类 k 的 $msdn$ 值越小说明该聚类越有可能是攻击聚类,检测算法如下:

(1)利用单维聚类算法对流数据集进行单维聚类,由于端口维和协议维层次结构较简单,而 IP 维层次结构较复杂,因此采用不同的聚类算法。对于端口维和协议维,直接扫描记录集即可,以协议维为例,算法如下:

```
for flow in flowset:
protcluster. volume += flow. protocol. bytes
protcluster. perc = protcluster. volume/totalvolume
```

(2)对于 IP 维,采用自底向上的算法构造聚类树。先扫描记录集,构造出叶子节点,对叶子节点按 IP 值从小到大排序。对于每个叶子节点,判断它与当前/31 前缀节点的关系,若为当前/31 前缀节点的孩子节点,则将该叶子节点流量加到/31 前缀节点流量上,并判断该节点是否为显著节点。否则新建一个/31 前缀节点,加入到/31 前缀节点链表中。依次建立每一层的树节点,直到/8 前缀节点,并对该聚类树进行压缩。这样就得到了一个 IP 聚类树。

(3)组合单维显著聚类得到流数据上的多维显著聚类。

①因为协议维、端口维的层次较少,首先对协议维和端口维进行聚合。分别取协议维、源端口维、目的端口维的根节点作为聚类树的根节点,然后分别取每一维的孩子节点与其他两维节点组合来派生第一层孩子节点,重复上述步骤直到叶子节点。

②压缩该聚类树。对于每一个节点,对它的孩子节点流量进行排序,找出最大值,若该节点流量减去孩子节点流量的最大值小于门限值,则将该节点压缩掉。

③将上述得到的协议维、端口维聚类树与 IP 维进行聚合。同样取每一维的根节点作为聚类树的根节点,然后依次派生每一层孩子节点。

④最后压缩多维聚类树,得到多维显著聚类。

(4)对于每一个多维显著聚类 k,计算 k 中每一个样本的距离 $d_k(i)$,即样本 x_i 到聚类 k 中其他样本 x_j 的最小欧氏距离。$d_k(i)=d(i,\arg_j\min d(i,j))$,$d(i,j)=\sqrt{(\sum_{i=1}^{p}(x_i-y_i)^2)}$ 为样本 x_i,x_j 的欧氏距离,每个样本为 p 维数据点,文中 $p=2$,分别为样本的时间戳和长度。

(5)计算聚类 k 中样本的平均距离

$\bar{d}_k=\frac{1}{m}\sum_{i=1}^{m}d_k(i)(x_i\in k)$。

(6)计算聚类 k 的距离标准差

$msd(k)=\sqrt{(\sum_{i=1}^{m}(d_k(i)-\bar{d}_k)^2)/m}(x_i\in k)$。

(7)计算聚类 k 的异常度 $msdn(k)=\frac{msd(k)}{N}$,并对 $msdn$ 值进行排序,若 $msdn$ 值明显偏小,则该聚类为攻击聚类。

4　实验结果

使用本算法实现了一个异常检测系统,并对其进行了测试。实验数据集采用了 MIT 林肯实验室提供的标准数据集 DARPA 2000 LL_DDoS_2.0.2,该数据集包含了目标为 131.84.1.31 的 DDoS 攻击的异常流量。参数 H,C 分别表示流量门限值和压缩门限值,分别对 H,C 取不同值进行实验,结果如表 1 ~ 表 3 所示(取前 5 个聚类结果):

表 1　DAPRA 数据集多维聚类结果($H=C=3\%$)

Cluster	perc	msdn
(*,131.84.1.31/32,T,high,high)	3.5%	0.00035
(*, 172.16.116.194/32, T, 80, high)	10.4%	0.61284
(*, 172.16.113.207/32, T, 80, high)	9.1%	0.72491
(*, 172.16.116.44/32, T, 80, high)	5.3%	1.02413
(*, 172.16.115.87/32, T, 80, high)	5.2%	1.14511

表 2　DAPRA 数据集多维聚类结果($H=C=6\%$)

Cluster	perc	msdn
(*, 172.16.116.194/32, T, 80, high)	10.4%	0.61284
(*, 172.16.113.207/32, T, 80, high)	9.1%	0.72491
(209.0.0.0/9, *, T, 80, high)	8.3%	0.73080
(*, 172.16.113.128/26, T, 80, high)	8.8%	1.67168
(*, 172.16.113.128/26, T, *, *)	11.4%	9.69072

表 3　DAPRA 数据集多维聚类结果($H=C=1\%$)

Cluster	perc	msdn
(*,131.84.1.31/32,T,high,high)	3.5%	0.00035
(172.16.112.100/32, 172.16.115.20/32, U, 1033, 53)	2.3%	0.00255
(*, 172.16.112.50/32, T, high, 23)	2.4%	0.42506
(*, 172.16.116.194/32, T, 80, high)	10.4%	0.61284
(*, 172.16.113.207/32, T, 80, high)	9.1%	0.72491

从表 1 的实验结果可以看出,多维规则(*,131.84.1.31/32,T,high,high)的流量占总流量的 3.5%,大于 $H=3\%$,说明该方法能有效的对网络中的流量进行多维聚类。并且观察它的 $msdn$ 值会发现,它的 $msdn$ 值较其他聚类明显偏小,因此该聚类为攻击聚类,说明该方法可以有效判断多维聚类的性质,实现异常检测。

不同的参数设置对结果有较大的影响,表 2 的实验中,设置参数 $H=C=6\%$,发现聚类结果中聚类数目显著减少,且没有攻击聚类,这是因为攻击聚类的流量百分比为 3.5% <6%,被压缩掉了。虽然没有得到攻击聚类,但是实验中发现系统运行时间有所减少。因此参数设置越大,就有可能漏掉攻击聚类,造成漏检,但是适当提高参数值,能提高检测效率。表 3 的实验中,设置参数 $H=C=1\%$,发现聚类结果中聚类数目显著增加,且有较多的叶子节点聚类,这是因为门限值越小,叶子节点越有可能形成聚类,并且在实验过程中发现系统运行的时间明显变长。因此参数设置越小,系统运行时间变长,检测效率降低。

5　结束语

异常流量与正常流量在流量特征上有明显的区别,可以通过分析流量特征来识别异常流量。文中提出了一种基于多维聚类分析的异常检测方法,能够自动对网络中的流量进行多维聚类,并对多维聚类进行分析,计算其异常度从而实现异常检测。下一步工作考虑继续完善本系统,并部署到实际的网络环境中进行测试。

参考文献:

[1] 严　芬,王佳佳,赵金凤,等. DDoS 攻击检测综述[J]. 计算机应用研究,2008,25(4):966-969.

[2] 胡振宇,辛　毅,方滨兴. 网络蠕虫检测方法的研究[J]. 微计算机信息,2008,24(2):64-65.

[3] 李和平,胡占义,吴毅红. 基于半监督学习的行为建模与异常检测[J]. 软件学报,2007,18(3):527-537.

[4] 杨新宇,侯光霞,杨树森. 带可信度评估的连续小波分布式拒绝服务攻击检测算法[J]. 西安交通大学学报,2008,42(8):936-939.

[5] Barford P, Kline J, Plonka D. A signal analysis of network traffic anomalies[C]//Proceedings of ACM SIGCOMM Internet Measurement Workshop. [s. l.]:[s. n.],2002:71-82.

[6] Kim S, Reddy A, Vannucci M. Detecting traffic anomalies through aggregate analysis of packet header data[C]//Proc of Networking 2004 (LNCS 3042). Berlin: Springer Verlag, 2004:1047-1059.

[7] Chhabra P, John A, Saran H. PISA: automatic extraction of traffic signatures [C]//Proc of Networking 2005 (LNCS 3462). Berlin: Springer Verlag, 2005:730-742.

[8] Estan C, Savage S, Varghese G. Automatically inferring patterns of resource consumption in network traffic[C]//Proc. of

(下转第 142 页)

则转到步骤(1);

(6)任意选择一点 $G' \in E(F_q)$,设置 $G=(N/n)G'$,重复这一步骤,至 $G \neq O$(为无穷远点)。

对于 $\#E(F_q)$ 的计算是一个纯数学问题。Rschool、Atkin、Elkies、Morain,Lercie 等人为此做了不少工作,Rschool 提出的著名 School 算法,经过 Atkin 和 Elkies 的改进提出了 SEA(School Elkies Atkin)算法。后来,Morain,Lercier 等专家又对 SEA 做了进一步的改进,目前,SEA 已被公认为是计算圆锥曲线的阶的比较有效的计算方法。除了 SEA,Satoh 还提出了 Satoh 算法以及目前对二进制域效率较高的 AGM 算法、MSS 算法、Satoh-FGH 算法、SSTT 算法等[8,9]。

2.5 验证域参数

在实际应用中,完全有可能会出现这样的问题:即无效的域参数的插入或者传输的错误,因此在使用域参数前必须对域参数来进行验证,以此保证域参数所须具备的数学特性,进而确保密码体制的安全。

在此给出对验证域参数的算法[9,10]:

输入的数据:$T=(q,a,b,G,n,h)$

输出的结果:T 有效或无效

(1)验证 q 为奇素数;

(2)验证 $G \neq O$;

(3)验证 $a,b,x_G,y_G \in F_q$;

(4)验证曲线是随机生成的;

(5)验证 a,b 是否满足曲线方程($4a^3+27b^2 \neq 0$);

(6)验证 G 是曲线上的一点;

(7)验证 n 是素数;

(8)验证 $n > 2^{160}$ 且 $n > 4\sqrt{q}$;

(9)验证 $nG=O$;

(10)验证 $h=\lfloor (\sqrt{q}+1)^2/n \rfloor, h=n$;

(11)验证对于每个 $k(1 \leqslant k \leqslant 20)$;

(12)验证 $n \neq q$;

(13)若上述有任一验证失败,那么 T 是无效的;如果验证通过则认为 T 是有效的。

综上论述,可以看到,域参数的生成是复杂的,所以在实际应用中大家可以选择 NIST 所推荐的安全曲线及参数,并将产生的参数放置于可信任的机构如 CA 中,需要时,从 CA 获得有效的参数,可由 CA 确保参数的有效性及安全性(CA:认证权威机构)[11]。

3 结束语

椭圆曲线密码体制的安全性是基于椭圆曲线离散对数的 NP 难解问题,而安全椭圆曲线的选取则是建立椭圆曲线密码体制的基石,所以曲线的安全是保证密码体系安全的重要因素。在过去的十多年里,椭圆曲线离散对数问题受到了数学界的极大关注[12]。目前,还没有发现椭圆曲线离散对数(ECDLP)有哪些特别大的弱点。

参考文献:

[1] 杨 剑,杨铭熙,李腊元. 增强安全的 IEEE802.15.4 协议研究[J]. 计算机技术与发展,2007,17(12):136-139.

[2] 张晓丰,樊启华,程红斌. 密码算法研究[J]. 计算机技术与发展,2006,16(2):179-180.

[3] 于雪燕,胡金初,柴春铁. 椭圆曲线密码体制及其参数生成的研究[J]. 计算机技术与发展,2006,16(11):160-161.

[4] 张 雁,林 英,郝 林. 椭圆曲线密码体制的研究热点综述[J]. 计算机工程,2004(2):127-129.

[5] 刘志猛,彭代渊. 基于椭圆曲线加密体制的实现[J]. 信息安全与通信保密,2006(4):94-96.

[6] 张龙军,沈钧毅,赵 霖. 椭圆曲线密码体制体制性研究[J]. 西安交通大学学报,2001(10):1038-1041.

[7] 王衍波,薛通编. 应用密码学[M]. 北京:机械工业出版社,2003.

[8] IEEE P1363/D6(Drafe Version 6). Standard Specification for Public Key Cryptography [EB/OL]. 2004. http://grouper.ieee.org/groups/1361/P1363/draft.html.

[9] 张 雁,林 英,郝 林. 构建安全椭圆曲线密码体制的关键问题[J]. 计算机应用,2004(12):82-84.

[10] Public Key Cryptography for the Financial Services Industry: The Elliptic Curve Digital Signature Algorithm (ECDSA) [S]. 1998.

[11] Johnson D, Menezes A. The Elliptic Curve Digital Signature Algorithm (ECDSA) [R]. Canada: University of Waterloo, 2000.

[12] 王平水,杨桂元. 基于有限域上圆锥曲线的公钥密码系统[J]. 微机发展(现更名:计算机技术与发展),2005,15(6):99-101.

(上接第 139 页)

ACM SIGCOMM Conference. [s.l.]:[s.n.],2003.

[9] Yasami Y, Mozaffari S P. A novel unsupervised classification approach for network anomaly detection by k-means clustering and ID3 decision tree learning method[J]. ACM Journal of Supercomputing,2010,53(1):231-245.

[10] Park N H, Oh S H, Lee W S. Anomaly intrusion detection by clustering transactional audit streams in a host computer[J]. Information Sciences,2010,180(12):2375-2389.

[11] 李 娜,钟 诚. 基于划分和凝聚层次聚类的无监督异常检测[J]. 计算机工程,2008,34(2):120-123.

[12] 周亚建,徐 晨,李继国. 基于改进 CURE 聚类算法的无监督异常检测方法[J]. 通信学报,2010,31(7):19-23.