基于服务特征分析与统计的入侵检测技术

孟 杰,陈 行,薛 俊,陶 军
(东南大学 计算机科学与工程学院,江苏 南京 210096;
东南大学 计算机网络与信息集成教育部重点实验室,江苏 南京 210096)

摘 要:提出一种新颖的基于服务特征分析的入侵检测方法。在处理网络审计数据时,首先针对网络服务进行特征分析,将审计数据按照网络应用进行区分,然后使用统计方差模型对应用区分后的审计数据进行检测;另外,在传统的统计方差模型基础上,提出加权的方法调整可信区间,提高检测率。选用KDDCup 1999 Data网络连接数据集进行实验,基于服务特征分析与统计的入侵检测方法在不增加虚警率的情况下,可以得到更高的检测精度。结果说明,该方法是行之有效的。

关键词:入侵检测;统计;服务特征分析;应用区分
中图分类号:TP393.08 **文献标识码**:A **文章编号**:1673-629X(2009)09-0146-04

Intrusion Detection Technology Based on Service Characteristics and Statistics

MENG Jie, CHEN Hang, XUE Jun, TAO Jun
(School of Computer Science & Engineering, Southeast University, Nanjing 210096, China;
Ministry of Edu. Key Lab. of Computer Network & Info. Integration, Southeast University, Nanjing 210096, China)

Abstract: A new method for intrusion detection is proposed based on network service characteristics. While processing the network audit data, do the characteristics analysis with network services and differentiate the audit data by network applications. Then finish the detection with statistics variance model. What's more, while adopting the traditional statistics variance model change the credibility interval by weight to improve the performance. At last, use the data set of KDDCup'99, and the result of experiment shows the proposed method could achieve better detecting rates.

Key words: intrusion detection; statistics; service characteristic analysis; application differentiation

0 引 言

随着互联网的发展,整个世界经济正在迅速地融为一体,计算机网络已经成为国家的经济基础和命脉,但是伴随着网络的发展,也产生了各种各样的问题,其中安全问题尤为突出。传统上,一般采用防火墙和杀毒软件进行安全保护。而随着攻击者手段的日趋成熟,攻击工具与手法的日趋复杂多样,单纯的防火墙加杀毒软件的策略已经无法满足对安全高度敏感的部门的需要,网络的防卫必须采用一种纵深的、多样的手段,因此,就产生了以入侵检测技术为主的主动防护技术[1],它在计算机安全防护体系结构方面起到了非常有意义的作用。入侵检测技术可以分为两大类[2]:一种是误用检测方法,已知的每种入侵方法都表示成一条入侵规则,将采集到的信息与入侵规则集,进行匹配确定入侵行为;另一种称为异常检测方法,它根据系统历史活动记录,为每个系统或用户建立正常活动的模型,通过比较当前采集的信息与正常活动模型之间的差异标记出异常信息。

目前,异常检测是入侵检测研究的主要方向[3],其特点是不需要过多的有关系统缺陷的知识,具有较强的适应性,并且能够检测出未知的入侵模式。异常检测的关键问题在于正常行为模式的建立以及如何利用正常行为模式对当前行为进行比较和判断。通常用到的异常检测方法有[4]:统计方法、免疫学方法、神经网络、数据挖掘[5,6]、机器学习[7]等等。经过多年的研究发展,已经取得了很好的检测效果。但是面对网络技术的快速发展和攻击行为的日益复杂,还有很多的不

收稿日期:2009-01-07;**修回日期**:2009-03-25
基金项目:国家自然科学基金重大研究计划项目(90604003);国家自然科学基金资助项目(60603067)
作者简介:孟 杰(1982-),男,江苏盐城人,硕士研究生,研究方向为网络安全、机器学习、入侵检测;导师:吴国新,教授,博士生导师,研究方向为网络安全和网络可信技术、高性能网络、基于P2P的资源共享技术。

足,如检测率低而误警率高,系统资源占用大而效率低等。

笔者在传统的统计方差模型的基础上,引入了服务特征分析来规整网络审计数据,并对可信区间范围进行加权,提高检测效率。首先针对网络服务进行特征分析,将审计数据按照网络应用进行划分,然后使用先验数据和统计的方差模型求出均值和标准偏差,从而得到可信任的区间范围,并使用加权的方法调整区间,最后,将该可信区间应用于入侵检测中,如果当前用户的行为超出了可信区间,则标示异常。

1　基于统计的异常检测算法

1.1　统计方法介绍

统计分析常用在基于异常的入侵检测系统中,是一种比较成熟的入侵检测方法。它使入侵检测系统能够学习主体的日常行为,将那些与正常活动之间存在较大统计偏差的活动标识成异常活动。它的主要优点是可以"学习"用户的使用习惯,从而具有较高检测率和可用性,所采用的技术和方法在统计学中已经得到很好的研究,例如:位于标准方差两侧的数据可以认为是异常的;主要缺点是异常阈值较难确定,阈值设置过高或过低均会导致误报事件。

统计模型在入侵检测中的应用历史可以追溯到技术发展的早期阶段,其应用非常广泛,效果也很好,例如经典的 IDES 统计异常检测引擎,以及后继的 NIDES 系统和 EMERALD 系统等。Denning 提出了用于异常检测的 5 种统计模型,其中第二种是方差模型[8],该模型成立的一个假设基础是系统当前状态特征可以采用数据的均值和标准偏差两个度量参数来刻画。在检测过程中,如果用户行为超出了可信任的区间范围,则标识为异常行为。

文中主要关注服务特征分析对入侵检测系统的作用,因此检测算法采用经典的统计方差模型,并使用加权的方法调整可信区间。

1.2　应用于 Web 服务的方差模型

这里使用的检测属性为 src_bytes(网络连接源端到目的端的数据字节数),以随机变量 X 表示。根据中心极限定理,如果所研究的随机变量 X 可以表示成很多个独立的随机变量 $X_1, X_2, \cdots, X_n$ 之和,只要每个 $X_i (i = 1,2,\cdots,n)$ 对 X 只起微小的作用,不管这些 X_i 服从什么分布,在 n 比较大的情况下,就可以认为 X 服从正态分布。由于文中采用的异常检测属性是独立的随机变量,因此可以使用该定理进行估计。

假设每个连接字节数为 t_i,前 n 个连接数据字节数总长度为 T_n,平均长度为 $\overline{T}_n$。则有

$$\overline{T}_n = T_n/n = \sum_{i=1}^{n} t_i/n \tag{1}$$

标准方差是测量数据的偏差,如果数据离平均值近,则置信区间较窄,对于 n 个数据值,样本标准差对总体标准差的无偏估计定义为:

$$s_n = \sqrt{\frac{\sum_{i=1}^{n} t^2 - nT_n^2}{n-1}} \tag{2}$$

样本均值和标准差能为检测属性(如:源端到目的端的数据字节数)的总体均值构造一个置信区间。样本均值的标准偏差为:

$$s_{\overline{x}_n} = s_n/\sqrt{n} \tag{3}$$

因此总体均值置信度为$(1-\partial)$的置信区间为:

$$(\overline{T}_n - \frac{s_n}{\sqrt{n}} z_{\frac{\partial}{2}}, \overline{T}_n + \frac{s_n}{\sqrt{n}} z_{\frac{\partial}{2}}) \tag{4}$$

$z_{\frac{\partial}{2}}$ 是一个标准正态分布的分位数,可以从正态分布表获取。样本中元素数目 n 越大,样本均值的偏差越小,其总体均值的偏差也就越小,如果当前网络连接从源端到目的端的字节数满足式(4)的要求,说明当前行为正常,如果测度不满足式(4)的要求,则说明当前行为异常,发出警报。

2　服务特征分析和加权

2.1　网络服务的特征分析

如何减少入侵检测系统的漏报和误报,提高其安全性和准确度是入侵检测系统的研究核心。从入侵检测系统的组成(见图 1)来看,可以从审计数据和检测算法来考虑,文中对审计数据做了进一步的研究。入侵检测系统发展到现在,已经出现了各种各样的数据源,包括系统日志、资源消耗(CPU、I/O 和内存使用等)、进程命令、网络数据等等,这里主要研究网络入侵,数据源是采集到的网络数据。

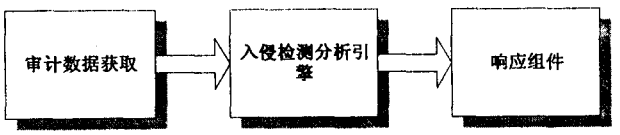

图 1　入侵检测系统组成

网络数据很多也很复杂,它是由网络服务(与端口对应)产生的,例如:Web 应用、FTP 应用、Email 应用等等。各种网络服务具有不同的作用和特点[9],例如:Web应用主要基于 http 协议(对应 80 端口),用于信息的发布和浏览,通常人们浏览的网址很多,安全程度也不一样。FTP 应用基于 FTP 协议(对应 21 端口),主要用于文件的共享,人们使用的 FTP 地址较为固定,数据传输在一段时间内会较为密集。Email 应用主

要基于 SMTP 协议和 POP3 协议，主要用于消息传递，人们拥有的邮箱一般都是固定的，这就意味着邮箱服务器的地址和数目相对固定。各种应用对应于不同的协议、工作机制和消息格式，面临的攻击威胁也是不一样的。这些都导致了网络应用产生的数据具有各自的特点。

这就引导我们采用分而治之的策略，将网络数据按照应用不同进行区分，从而减小审计数据的规模，提高检测速度，也使得审计数据更加规整，有利于做进一步的统计分析。主要将网络数据按照 Web 应用、FTP 应用、Email 应用、telnet 应用、其他应用总计 5 种进行区分，再将应用区分后的审计数据使用统计的方差模型做进一步的检测。

2.2 异常检测系统的属性选择

网络数据有很多可供选择的检测属性，如：网络报文的各个字段，帧(目的 mac 地址，源 mac 地址)，IP 报文(源 IP 地址，目的 IP 地址，服务类型，报文总长度，标识符，标志，生命周期)，TCP 报文(源端口，目的端口，序列码，窗口)，UDP 报文(源端口，目的端口)等等。各个检测属性在不同的应用中具有不同的特点(均值和标准偏差不同)，选择属性以及处理属性间的关系对于检测系统的性能影响很大。

定义 1 可测度集。设 F 为一组属性值集合，包含所有审计数据或连接记录中可能出现的属性值，例如源地址、目的地址、报文长度、端口号等，称之为可测度集。设 $V = \{\text{attribute1}, \text{attribute2}, \ldots, \text{attribute}n\}$ 为 F 的子集，称之为记录属性。可测度集表明一个数据包或一条连接的属性集合，将需要检测的信息作为记录属性。

文中使用的是 KDDCup'99[10]数据集，它提供了 41 个检测属性，这些属性都是针对攻击选出来的，具有很有的检测效果，因此直接使用该数据集提供的属性，并重点关注了其中的 9 个连续属性，则有可测度集 $V = \{$src_bytes, dst_bytes, count, srv_count, srv_diff_host_rate, dst_host_count, dst_host_srv_count, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate$\}$，在系统实现的过程中，发现其中部分属性的统计效果很差，文中选择了含义较为简单的、统计效果较好的 src_bytes(网络连接源端到目的端的数据字节数)作为实验中检测属性，重点关注服务特征分析给检测系统带来的影响。

2.3 可信区间的加权改进

统计方法中的阈值确定一直是统计方法的难题，太小的值会产生大量的误报，太大的值会产生大量的漏报。因此如何提高监测的准确度是必须要解决的一个问题。在实验的过程中，发现式(4)的信任区间，检测的效果并不好，检测率和误报率都相当高，经分析是信任区间太小导致了大量的误报，因此引入了加权因子 θ，来改变区间的大小，进一步提高检测的性能，则有置信度为$(1-\partial)$的置信区间为：

$$\left(\bar{T}_n - \frac{s_n}{\sqrt{n}} z_{\frac{\partial}{2}} \theta, \bar{T}_n + \frac{s_n}{\sqrt{n}} z_{\frac{\partial}{2}} \theta\right) \tag{5}$$

3 实验和讨论

3.1 数据集描述和处理

文中使用 KDDCup'99 训练数据集，它是由 MIT 林肯实验室公开提供的 DARPA 入侵检测评价计划中的网络通信数据集经处理后形成的。该数据集文件由大量的网络协议连接记录构成，约有 5000000 条。文中使用的是它提供的一个 10% 的训练数据子集，该子集包含有 494020 条记录，下面是部分连接记录的例子(见图 2)。

```
0, tcp, smtp, SF, 1000, 494, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 128, 244, 0.95, 0.02, 0.01, 0.01, 0.00, 0.00, 0.00, 0.00, normal.
1, tcp, smtp, SF, 950, 493, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1, 0.00, 0.00, 0.00, 0.00, 0.50, 1.00, 0.00, 138, 213, 0.95, 0.01, 0.01, 0.01, 0.00, 0.00, 0.00, 0.00, normal.
0, icmp, ecr_i, SF, 1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 316, 316, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 148, 3, 0.02, 0.02, 0.02, 0.00, 0.00, 0.00, 0.00, 0.00, smurf.
0, icmp, ecr_i, SF, 1032, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 511, 511, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 158, 13, 0.08, 0.02, 0.08, 0.00, 0.00, 0.00, 0.00, 0.00, smurf.
```

图 2 数据集连接记录示例

每条记录有 42 个字段，42 个字段中有 4 个字符串型字段，其它为数值字段。其中第 3 个字段表示连接记录是哪种网络应用，总计有 66 种，第 42 个字段是类别标记，标明是正常还是入侵记录，如果是入侵则标明是哪一种入侵，入侵的种类总计有 22 种，可以分为 4 类，分别为 R2L 攻击，U2R 攻击，DoS 攻击，probing 攻击。

文中将该 10% 的训练数据子集按照网络应用(Web 应用、FTP 应用、Email 应用、telnet 应用、其他应用)的种类划分，使用未区分应用的数据子集和 Web 应用(只包括 http 服务，未包含 https 服务)的数据子集进行实验，其汇总情况见表 1。然后通过随机采样的方法得到 10000 条正常记录的训练数据集和分别为 2500 条，5000 条，10000 条，20000 条，40000 条的测试数据集。

依据表中的数据可以看出网络数据集中各种应用的总记录数和易遭受攻击的记录数存在较大的差异，因此需要区分应用来对审计数据进行预处理。

文中使用的检测标准有：检测率 TPR(True Positive Rate)为攻击报文的检测比率，虚警率 FNR(False Negative Rate)为正常报文的误报比率。

表 1　区分应用的数据子集汇总情况

数据集	Normal	R2L 攻击	U2R 攻击	DoS 攻击	probing 攻击	总计
未区分应用的数据集	97277	1131	52	391453	4107	494020
Web 应用的数据集	61886	4	0	2395	8	64293
ftp 应用的数据集	373	313	3	104	5	798

3.2　加权因子的影响

采用 10000 条正常记录的训练数据和 20000 条的测试数据做加权因子的实验,置信度为 99%,加权因子从 1 至 300,结果见图 3。

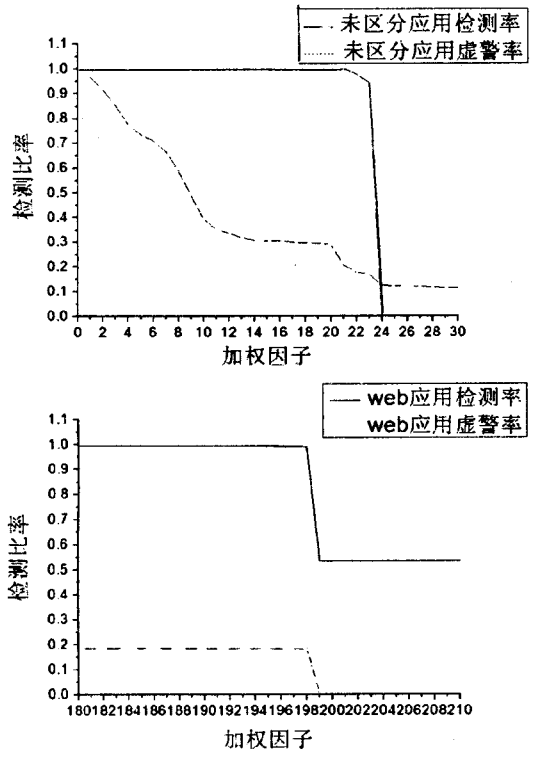

图 3　加权因子对检测性能的影响

图 3 描述了加权因子对检测性能的影响。随着加权因子增大,虚警率会减小,检测率也相应减小。这是因为加权因子增大,可信区间就变大,因此正常事件的误报减小,但同时更多的攻击事件掉入可信区间,检测率降低。另外,随着加权因子增大,可信区间变大,Web 应用的检测性能降低较为平滑,未区分应用的检测性能降低曲线则过于曲折,这也正说明了服务特征分析和区分应用可以使审计数据更加规整。实验结果说明,统计的方差模型的信任区间大小对于检测性能影响较大,不同的加权因子会有不同的检测效率,选择一个较好的加权因子能够显著提高检测系统的性能。

3.3　区分应用的影响

采用 10000 条正常记录的训练数据和 2500 条,5000 条,10000 条,20000 条,40000 条的测试数据集进行区分应用的影响实验,置信度为 99%,未区分应用数据集的加权因子为 23,Web 应用的加权因子为 197。结果见图 4。

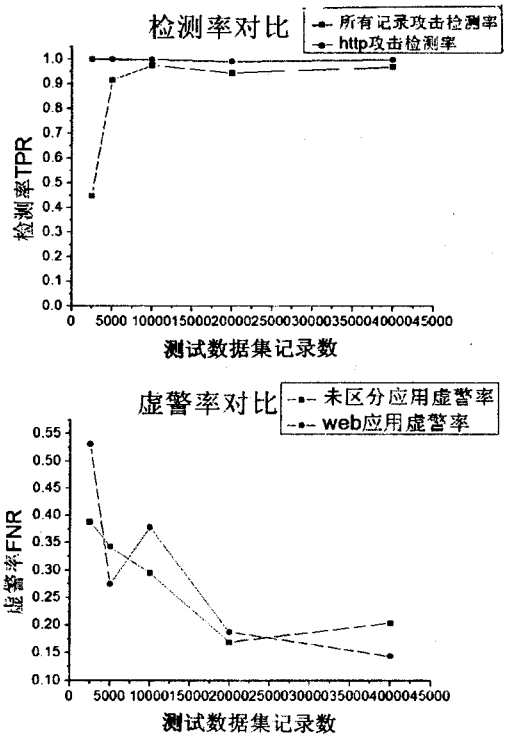

图 4　区分应用与未区分应用的对比

图 4 描述了区分应用与未区分应用对检测性能的影响。在不同大小的测试数据集下,Web 应用和未区分应用的虚警率基本相持,而检测率则是 Web 应用明显占优。这是因为经过服务特征分析和应用区分后,审计数据更加规整,因此会有更好的检测效果。另外,在检测率对比上,Web 应用的检测率对于测试数据集的大小不太敏感,这也在一定程度上说明,服务特征分析和应用区分可以使入侵检测系统具有更好的通用行和适应性。实验结果说明,Web 引用在不增加虚警率的情况下,可以显著提高攻击报文的检测率,即服务特征分析和区分应用有利于检测性能的提高。

4　结束语

文中提出了应用区分的概念,减少和规范审计数据,并在传统的统计方差模型基础上,引入了加权因子来调整可信区间。实验结果表明,与传统的统计方法相比,在虚警率基本相持的情况下,该方法具有更好的检测性能。但是如果一个攻击涉及到多个网络应用,

(下转第 167 页)

结果大多略有偏大,该误差产生原因主要有两点:一是对压力传感器零点漂移的抑制不够理想;二是由于利用示波法测量血压,特征点的确定只能依赖采集样本的统计归纳,有一定的离散性,因此会存在一定的误差。

表 1　测试结果一

(收缩压、舒张压:mmHg,心率:次/分)

测量者	本电子血压计			水银血压计		
	心率	舒张压	收缩压	心率	舒张压	收缩压
一	64	70	109	63	68	107
二	72	69	113	70	70	112
三	70	54	104	69	53	103

表 2　测试结果二

(收缩压、舒张压:mmHg,心率:次/分)

测量者	本电子血压计			水银血压计		
	心率	舒张压	收缩压	心率	舒张压	收缩压
一	65	70	110	64	65	108
二	70	69	112	69	71	113
三	68	56	105	69	57	106

表 3　测试结果三

(收缩压、舒张压:mmHg,心率:次/分)

测量者	本电子血压计			水银血压计		
	心率	舒张压	收缩压	心率	舒张压	收缩压
一	66	71	109	65	68	107
二	72	71	113	71	70	112
三	68	55	107	68	54	105

另外,实验中一次测量时间 20～25 秒,测量时间短。这是因为该血压计采用充气过程测量,放气速度很快,因此测量时间更短,且对同一测量者的多组测量结果看出,该血压计测量的重复性和一致性好,特别是对不同身体状况的测量者具有良好的个体适应性。大屏液晶显示提供了良好的人机界面,系统低功耗使得系统耗电量小,因此血压计可在临床诊断和家庭医疗保健中应用。

5　结束语

文中给出了一种基于模数混合的 Fusion FPGA 的电子血压计的完整设计方案,该方案硬件集成度高,充分利用了 Fusion FPGA 模数混合的特点和丰富的片上资源,减少了电路板面积和系统总成本,对传统的幅度系数法进行了改进。实验证明,这种电子血压计测量精度高、存储容量大、功能丰富、功耗低、使用方便。且 Fusion FPGA 丰富的逻辑资源和内部集成的多种功能部件使得系统可很方便进行升级或功能扩展,可在此基础上设计多功能动态生理信号监测仪,在临床医疗上推广使用。

参考文献:

[1] 钱　峰,刘　晰.基于示波法的电子血压计实现[J].仪器仪表学报,2006,27(6):1534－1535.

[2] Ball－Llovera,Del Rey A,Ruso R,et al. An Experience in Implementing the Oscillometric Algorithm for the Non－Invasive Determination of Human Blood Pressure[J]. IEEE Engineering in Medicine and Biology － Proceedings,2003(4):3173－3175.

[3] Kim T K,Chee Y J,Lee J S,et al. A new blood pressure measurement using dual－cuffs[J]. Computers in Cardiology,2008,35:165－168.

[4] 包旭鹤.便携式电子血压计设计[J].现代电子技术,2007(8):7－10.

[5] 刘宝华.一种新型电子血压计的研制[J].燕山大学学报,2005,29(1):60－63.

[6] 田辉勇,苏永春,刘文军,等.基于充气测量的电子血压计研制[J].医疗卫生装备,2007,28(8):16－18.

[7] 周立功.Actel FPGA 原理与应用[M].广州:广州致远电子有限公司,2007.

[8] 杜凌云,黄士坦.高速 A/D 与微处理器间的数据缓存技术[J].计算机技术与发展,2007,17(4):167－170.

(上接第 149 页)

文中的方法还有所欠缺,这也是下一步研究的内容。

参考文献:

[1] 蒋建春,马恒太,任党恩,等.网络安全入侵检测:研究综述[J].软件学报,2000,11(11):1460－1466.

[2] 肖竟华,卢　娜.基于网络的入侵检测系统的研究及实现[J].计算机技术与发展,2007,17(2):242－244.

[3] Kemmerer R A,Vigna G. Intrusion Detection:A Brief History and Overview[J]. Supplement to IEEE Computer (IEEE Security & Privacy),2002,35(4):27－30.

[4] 唐正军,李建华.入侵检测技术[M].北京:清华大学出版社,2004.

[5] 叶和平,尚　敏.一种面向入侵检测的数据挖掘算法研究[J].计算机技术与发展,2008,18(11):149－151.

[6] 罗军生,李永忠,杜　晓.基于模糊 C－均值聚类算法的入侵检测[J].计算机技术与发展,2008,18(1):178－180.

[7] 柏海滨,李　俊.基于支持向量机的入侵检测系统的研究[J].计算机技术与发展,2008,18(4):137－139.

[8] 程　光,龚　俭,丁　伟.基于抽样测量的高速网络实时异常检测模型[J].软件学报,2003,14(3):594－599.

[9] 中国互联网络信息中心.中国互联网络发展状况统计报告[EB/OL]. 2008. http://www.cnnic.net.cn/index/0E/00/11/index.htm.

[10] Stolfo S J,Fan Wei,Lee Wenke,et al. Task description of Kddcup'99[EB/OL]. 1999. http://kdd.ics.uci.edu/databases/kddcup99/task.html.