基于密度聚类算法的入侵检测研究

蔡伟鸿　刘　震
(广东汕头大学工学院计算机系,汕头 505063)
E-mail:Suada@eyou.com

摘　要　本文联系异常检测和数据挖掘，从理论上着重分析了在入侵检测系统中应用基于密度聚类算法的必要性和有效性,从 TCPDump 网络数据和系统日志中提取分析后生成特征数据,通过 Clenmine 中 CEMI 实现定制的基于密度的改进 DBSCAN 算法进行测试,结果表明利用该算法可以较好地识别分布式拒绝服务攻击等多种入侵行为。

关键词　异常检测　基于密度的聚类　数据挖掘

文章编号 1002-8331-(2005)21-0149-03　**文献标识码** A　**中图分类号** TP393.08

Intrusion Detection Research Based on the Density Clutering Method

Cai Weihong　Liu Zhen
(Department of Computer,College of Engineering,Shantou University,Shantou,Guangdong 505063)

Abstract: In this paper,we have discussed the anomaly detection with the data mining.And theoretically we have analyzed the feasibility and necessity of using the density clustering method in the intrusion detection system.Using the specific CEMI interface offered by the Clenmine,we can implement the improved DBSCAN method.With the characteristic data distilling from a mass of the network TCPDump and system log,we have done a experiment.The result testified the intrusion detection system based on the density clustering method can identify the Ddos detection effectively.

Keywords: anomaly detection,density clustering,Data Mining

由于计算机网络的复杂性,仅仅使用由防火墙和各种数据加密，安全认证等基于密码学的方法无法有效保证系统安全,入侵检测越来越显示出重要的作用。

1　引言

入侵检测技术主要包括:异常检测和滥用检测。异常检测是指将用户正常的习惯行为特征存储在数据库中,然后将用户当前的行为特征与特征数据库中的特征进行比较,如果两者的偏差足够大,则说明发生了异常。这种方法可以检测出未知的攻击类型,缺点是误检率比较高。滥用检测是指将已知攻击方式以某种形式存储在知识库中,然后通过判断知识库中的入侵模式是否出现来检测,如果出现,则说明发生入侵。这种方法只能检测已知攻击类型,但准确率比较高。而数据挖掘方法的优势在于它能从大量数据中提取感兴趣的、事先未知的知识和规律,而不依赖经验。在入侵检测中应用数据挖掘,可以从大量的审计数据中发现有助于检测的知识和规则,自动发现新的异常模式,从而实现对新的未知攻击模式的检测。数据挖掘算法通常可以分为关联分析,序列分析,分类分析和聚类分析。目前数据挖掘在 IDS 商业系统和理论研究主要集中在关联分析和序列分析。

关联分析:令 $I=\{i_1,i_2,\ldots,i_n\}$为项目集,D 是事务数据库,其中每个事务 T 是一个项目子集,并具有一个唯一的标识符 ID。关联分析是形如 $X\rightarrow Y$ 的蕴涵式,解释为“满足 X 中条件的数据库元组多半也满足 Y 中条件”。它有两个重要属性:支持度(support)和可信度(confidence),支持度是指包含 X 和 Y 的元组数与所有元组数的比值，可信度是指包含 X 和 Y 的元组数与包含 X 的元组数之比。

序列分析:发现数据之间的前后出现的因果关系。运用序列分析找出入侵行为的序列关系,从中可以提取出入侵行为之间的时间序列特征。

而聚类是观察式学习,是发现隐含于混杂数据对象的分类规则。作为统计学的分支,聚类分析的应用和研究,主要集中在基于距离的聚类分析，常用的主流数据挖掘工具 S-Plus,SPSS 以及 SAS 都支持该类聚类算法的实现。Portnoy 提出基于聚类分析的入侵检测方法，无监督异常检测算法(unsupervised anomaly detection algorithm)[2]，通过对未标识数据进行训练检测入侵。算法将数据实例进行正规化处理转换成为标准形式,采用标准欧几里德度量,采用单链法聚类,经过标识,分类检测入侵。但它需要假设给定的数据集服从一个随机分布(如正态分布等)，但实际的数据往往不符合任何一种理想状态的数学分布;同时该算法不适用于检测拒绝服务攻击等恶意攻击行为。

基于密度聚类的基本思想是大部分数据是正常的,正常数据会聚集在一起成为一个高密度簇。入侵数据差别很大并且数据很少。聚类后入侵数据会成为一个低密度簇。因此,先进行聚类分析，小簇数据即判定为入侵发生。由于基于密度聚类的 DBSCAN 算法[3]具有可以挖掘任意形状的聚类,并且对数据输入顺序不敏感,具有处理异常数据(噪音)等优点,我们考虑将 DBSCAN 算法加以改进用于检测 Ddos 等入侵攻击。

2　改进的 DBSCAN 算法

作者简介:蔡伟鸿,男(汉族),副教授,研究方向为计算机网络应用技术。刘震,硕士生,主要研究方向:网络安全。

2.1 算法中使用的概念

定义 1 ε 邻域：给定对象半径 ε 内的区域称为该对象的 ε 邻域。

定义 2 核心对象：如果一个对象的 ε 邻域至少包含最少数目的 MinPts 个对象，则称该对象是核心对象。

定义 3 直接密度可达：给定一个对象集合 D，如果 p 是在 q 的 ε 邻域内，而 q 是一个核心对象，则称对象 p 是从对象 q 出发是直接密度可达的。

定义 4 密度可达：如果存在一个对象链 $P_1,P_2,\dots,P_n,P_1=q$，$P_n=p$，对于 $P_i\in D,(1<i<n),P_{i+1}$ 是从 P_i 关于 ε 和 MinPts 直接密度可达的，则对象 P 是从对象 q 关于 ε 和 MinPts 密度可达。

定义 5 密度相连：如果对象集合存在一个对象 o，使得对象 p 和 q 是从 o 关于 ε 和 MinPts 密度可达的。则称对象 p 和 q 是关于 ε 和 MinPts 密度相连的。一个基于密度的簇是基于密度可达性的最大的密度相连对象的集合。

定义 6 核心距离(core-distance)[4]：一个对象 p 的核心距离是使得 p 成为核心对象的最小 ε。如果 p 不是核心对象，p 的核心对象没有定义。

2.2 算法描述

```
clusterNo=0;//初始化,簇的初始数目为 0
while(数据库非空)
{
    read(点 p)
    if (点 p 的 ε 邻域包含多于 MinPts 个点)
    {
        markPAsCore(p);//将 p 标识为核心对象,并将从这个核心对象直接密度可达的对象进行标识;
        clusterNo++;
    }
    if(clusterNo≥3)
    {
        compareDistance(p1,p2);//比较各簇核心对象的核心距离,选取较小的两个核心对象标记为 co1,co2;
        long distance=measureDistance(co1,co2);//将两个核心对象间的距离标记为 distance
        deviseNewArea();//设定以 co1,co2 的中点为圆心,以 1/2*distance 为半径的圆形区域为收敛域,
        chooseNewCore();//选取 ε 邻域内直接密度可达对象最多的点 p 为新簇的核心对象,并确定归并后 p 的核心距离;
        clusterNo--;
    }
}
```

2.3 算法实现

利用 Tcpdump 获取本机及局域网内进出的数据，经过预处理，将其转换成 ASCII 格式，得到有报文头和数据载荷的网络连接记录。经过进一步的信息提取和简单计算，可以获得每条连接记录的摘要信息，生成新的连接记录集和特征记录。取下列属性用于数据挖掘：

duration，protocol，wrong_fragment，SYNerror_no，REJerror_no，byte_from，byte_to，E_state，differentFlag，Pcercent_samesource，far_port，local_port，Rin，Rout，ΔR_state。

各个属性的含义分别为：

duration：连接持续的时间；

protocol：连接双方使用的协议。其取值有 smtp，http，telnet，ftp 等；

wrong_fragment：错误的分片数目；

SYNerror_no：具有“SYN”错误的连接所占的百分比；

REJerror_no：具有“REJ”错误的连接所占的百分比；

byte_from：表示提出请求方发出的字节数；

byte_to：表示响应方发送的字节数；

E_state：表示连接结束的状态，正常结束为 TRUE，非正常结束为 FALSE；

differentFlag：如果源地址/端口目的地址/端口，则为 1，反之为 0；

Pcercent_samesource 相同源主机连接所占的百分比；

far_port：表示远程主机的端口。例如常见服务的端口包括 FTP 服务器的端口号是 21，每个 TELNET 服务器的 TCP 端口号是 23，每个 TFTP 服务器的端口号是 69；

local_port：表示本地主机的端口号；

Rin：表示进入本地主机的数据流字节数；

Rout：表示从本地主机发出的数据流字节数；

ΔR_state：表示进入本地主机数据流字节数和从本地主机发出的数据流字节数的差额状态。

另外，我们采用来自主机 log 等的若干特征属性应用于数据挖掘。

FailureNo：表示 login 失败次数；

Percent_CPU：表示 CPU 利用率；

Percent_memory：表示内存利用率。

将上述属性分成三大类：

(1)区间标度类型变量(主要指可以连续度量的变量)

具体包括：连接持续时间 duration，错误的分片数目 wrong_fragment，具有“SYN”错误的连接所占的百分比 SYNerror_no，具有“REJ”错误的连接所占的百分比 REJerror_no，源主机发出的字节数 byte_from，相同源主机连接所占的百分比 Pcercent_samesource，目的主机发送的字节数 byte_to，表示进入本地主机的数据流字节数 Rin，远程主机的端口 far_port 和本地主机的端口 local_ip，login 失败次数 FailureNo，CPU 利用率 Percent_CPU，内存利用率 Percent_memory，表示从本地主机发出的数据流字节数 Rout。

(2)二元变量(即布尔型变量)

具体包括：是否正常结束状态 E_state，源地址/端口和目的地址/端口是否相等标记 differentFlag 入本地主机的数据流字节数和从本地主机发出的数据流字节数的差额标记 ΔR_state。当进入本地主机的数据流字节数和从本地主机发出的数据流字节数的差额大于 0 时，标记 ΔR_state 取 1，当该差额小于 0 时，取 0；当源地址/端口和目的地址/端口不同时，differentFlag 取 1，反之取 0。

另外，我们以 TCP 连接的终止过程为例，说明如何分析得到 E_state 值。假设主机 A 和主机 B 已经建立连接，现在 A 想终止和 B 建立的连接。正常结束过程如下：主机 A 首先向 B 发送一个 FIN 报文；主机 B 收到这个 FIN，发回一个 ACK，确认序号为 FIN 报文的序号加 1；主机 A 接收 ACK 报文；同时主机 B 还向上层应用程序传送一个文件结束符，服务器程序关闭这个连接，向主机 A 发送 FIN sep=y ACK 报文；主机 A 接收 B 发来的 FIN+ACK 报文；主机 A 向主机 B 发送确认 ACK 报文。在主机 A 上运行 TCPDump 得到原始数据，当数据包含发送 FIN，接收到 B 的 ACK 报文，接收到 B 的 FIN+ACK 报文段，发送的

ACK 报文段这四个完整的数据包时，我们认为 TCP 连接正常结束，取 E-state 为 1；反之，E_state 取 0。

(3)标称变量(是二元变量的推广，指具有多于两个状态的状态值的变量)

主要指双方使用的协议 protocol，试验中发现常见的协议有 smtp，http，telnet，ftp，finger，domain 等。

算法中采用的数据结构：采用相异度矩阵，表现形式是 $m\times m$ 维的矩阵。m 表示 m 个数据对象，矩阵中元素 $d(i,j)$ 表示对象 i 和对象 j 的相异度。

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ \cdots & & & \\ d(m,1) & d(m,2) & \cdots & 0 \end{bmatrix}$$

算法中用到的相异度数据的计算：

由于我们使用的数据集包含 3 种不同类型的变量，对象 i 和 j 之间的相异度定义为：

$$d(i,j)=p_1 d_{ij}^{(1)}+p_2 d_{ij}^{(2)}+p_3 d_{ij}^{(3)}$$

其中 p_1,p_2,p_3 分别是区间标度类型变量，二元类型变量，标称类型变量对应的权值，试验中分别取权值 p_1,p_2,p_3 为 0.1，0.7，0.2。$d_{ij}^{(1)},d_{ij}^{(2)},d_{ij}^{(3)}$ 分别是区间标度类型变量，二元类型变量和标称类型变量的相异度。

(1)区间标度变量

首先需要计算每个变量 S_i 的均值，$S_i=\frac{1}{n}(x_{1i}+x_{2i}+\ldots+x_{ni})$

$$d(i,j)=\frac{\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\ldots+(x_{ip}-x_{jp})^2}}{S_1+S_2+\ldots+S_p}$$

(2)二元变量(只有两个状态 0，1)

我们用 q 表示对象 i 和 j 值都为 1 的变量的数目，r 表示对于对象 i 值为 1 而对象 j 值为 0 的变量的数目，s 是对于对象 i 值为 0 而对象 j 值为 1 的变量的数目。

相异度采用 Jaccard 系数 $d(i,j)=\frac{r+s}{q+r+s}$ 度量。

(3)标称变量

我们用 p 表示全部变量的数目，m 表示取值相同的变量的数目。

$$d(i,j)=\frac{p-m}{m}$$

3 应用实例

我们采用了如图 1 所示的网络试验环境，共 4 台主机，分别分布在两个网段内，中间通过两台 Cisco2611XM 路由器 R1、R2 和 CiscoPIX-506E 防火墙连接。我们在主机 Host1 安装了分布式拒绝服务攻击工具 trino，trino 使用 master 控制主机 Host2 和主机 Host3，向主机 Host4 发动攻击，攻击从 Host4 使用 Tcpdump 开始收集数据 1 分钟后开始，持续时间为 20 秒。在主机 Host4 上使用 Tcpdump 收集 2 分钟的原始网络数据包的相关数据，经过分析整理得到以下相关数据如表 1 所示。

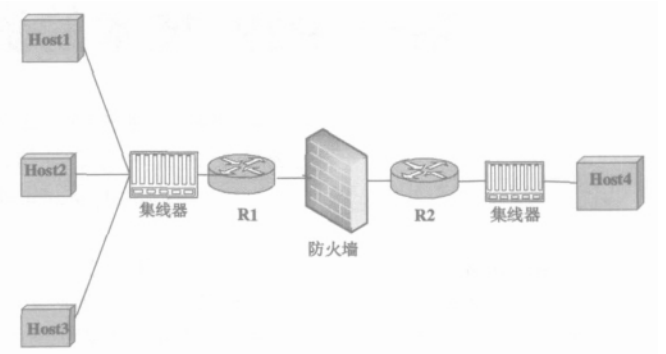

图 1 试验网络拓扑图

表 1 数据库中表的基本格式

duration	protocol	byte_from	byte_to	E_state	far_ip	far_port	...
...	...	...	...	...	...	...	...
2	telnet	68	182	TRUE	10.10.90.57	23	...
...	...	...	...	...	...	...	...

采用 Spass 公司的主流数据挖掘工具 Clenmine 中的 CEMI 定制实现上述的改进 DBSCAN 算法，我们可以观察到正常数据有 68%聚集在高密度簇中，异常的网络攻击数据有 76%聚集在低密簇中，从而说明使用基于密度的聚类挖掘算法可以很大程度上检测 DDos 等攻击。

算法存在的缺陷：需要根据网管人员的经验来确定输入参数 ε 和 MinPts 对对象进行聚类，具有较大的主观性和局限性，同时，由于算法需要反复进行多个簇的归并，重新确定归并后新簇的核心对象和核心距离，导致算法执行效率较差。另外，从理论上讲，关联序列分析、序列分析和聚类分析得出的规则存在出现冲突或者矛盾的可能性[5]，当规则出现冲突时，如何确定正确规则，过滤错误规则。这些都是我们下一步需要重点研究的问题。(收稿日期：2004 年 8 月)

参考文献

1.Lee W，Stolfo S J.Data Mining Approach for Intrusion Detection[C].In：Proceedings of the 7th Usen IX Security Symposium，San Antonio，TX，1998-01-26

2.Eleazar Eskin A A，Prerau m，Portnoy L et al.A geometric framework of unsupervised anomaly detection：Dectecting intrusions in unlabeled data.Data Mining for Security Applications(DMSA-2005)kluwer 2，2000

3.Jiawei Han，Micheline Kamber.Data Mining：Concepts and Techniques[M].Copyright 2001 by Morgan Kaufmann Publishers，Inc.242~243

4.M Ankerst，M Breunig，H-p Kriegerl et al.OPTICS：Ordering points to identify the clustering structure[C].In：Proc 1999 ACM_SIGMOD Int Conf Mangement of Data(SIGMOD'99)，Philadephia，PA，1999-06：49~60

5.A Berson，S J Smith，K Thearling.Building Data Mining Applications for CRM[M].New York：McGraw-Hill，1999