数据挖掘技术在蜜网中应用研究①

李巧君, 鲁华栋

(河南工业职业技术学院 计算机工程系, 南阳 473009)

摘 要: 部署蜜网(Honeynet)的目的之一就是收集数据，但若无法对捕获的数据进行分析处理，则该数据就毫无意义. 本文对蜜网中捕获的日志模块数据利用数据挖掘技术进行标记分类，使用分类算法对已经分好类的数据进行有规则的挖掘，从而发现入侵者的攻击方法，为未来各种攻击行为做好防御准备.

关键词: 数据挖掘; K-means 算法; 日志模块数据; Honeynet

Research on Data Mining in Honeynet

LI Qiao-Jun, LU Hua-Dong

(Computer Engineering, Henan Polytechnic Institute, Nanyang 473009, China)

Abstract: Collecting data areone of the aims of Honeynet, But how to analyze these collected data is the keypoint. Data mining is introduced to mark and sort the log module data prayed in Honeynet. And the sorted data were mined regularly with sort algorithm. By means of that, the attack method could be found, and good defense ways for varies attack manners would be deployed,Collecting data is one of the aims of Honeynet.

Key words: data mining; K-means algorithm; log module data; Honeynet

1 引言

蜜罐(Honeypot)[1]是一种专门设计成被扫描、攻击和入侵的网络资源，用于收集入侵者的访问信息，以日志的形式记录其访问活动，使入侵者在蜜罐中耗费精力和技术，从而保护正常的系统和资源. 蜜网(Honeynet)技术[2][3]是在蜜罐技术上发展起来的一个新的概念，又可称为诱捕网络，由一个或多个蜜罐组成. 对于蜜网络系统来说，网络交互产生的所有数据都是可以(而且是必须)被记录的，这些网络数据的寄存被称之为网络连接记录日志.

日志信息可以有效地应用于故障监控、对数据访问服务恶意攻击的调查，但若无法对日志数据进行分析，则其毫无用途，与日志记录相关的主要问题是海量数据的分析. 为了提高入侵检测任务的有效性，使用数据挖掘技术对网络日志数据进行分析.

数据挖掘[4][5]对数据进行挖掘分析，从海量的事务数据库中，发现支持度与置信度分别大于阈值的事务项之间相互关联的规则，帮助安全研究人员分析、学习攻击方法和攻击流程，以便更好地防御未来的攻击行为.

2 数据分析对象的选择和收集

本文的数据来源于一个基于某高校校园虚拟 Honeynet 中，该蜜网与 IDS、防火墙、路由器等技术结合使用，在该系统中一个蜜罐主机作为日志服务器收集攻击蜜网的日志信息. 蜜罐主机的日志信息主要来自于由 Sebek 客户端基于内核级的对于入侵者行为的记录，这些信息包括入侵者的击键序列和访问日志记录信息，远程服务器上的一个叫 snort 的数据库存储该数据[6]. 每条日志记录由一些固定的属性组成: 时间戳、协议、源 IP 地址、源端口、目的 IP 地址、目的端口、包长、操作系统等.

根据调查表明，大多数的网络攻击分析均是只需要协议头部信息和数据包实际内容的一部分，因此不记录数据包的全部，而是记录协议的头部信息以及实际内容的开头的前部分字节，作为详细的原始日志数

①收稿时间:2013-10-31;收到修改稿时间:2013-12-10

据，记录的数据包括网段内部主机之内，以及网段内部主机与外部主机之间通信的数据包及数据内容. 表 1 显示收集到的部分日志信息.

表 1 日志信息

Time	Hostname	Message
16:01:33	192.168.1.2	Killing attemptd connection: tcp [192.168.0.4:45673-192.168.1.3:350]
16:01:23	192.168.1.2	Connection dropped by reset: tcp[192.168.0.4:3402-192.168.1.3:130]
16:01:20	192.168.1.2	Connection dropped by reset: tcp[192.168.0.4:3400-192.168.1.3:33]
16:01:34	192.168.1.2	Connectionestablished: tcp[192.168.0.4:3245-192.168.1.3:33] <->sh scripts/unix/linux/qpop.sh 192.168.0.4:4350 192.168.1.3:33
16:01:35	192.168.1.2	Connectionestablished: tcp[192.168.0.4:5305-192.168.1.3:130] <->sh scripts/unix/linux/qpop.sh 192.168.0.4:5305 192.168.1.3:130
16:01:36	192.168.1.2	Connection dropped by reset: tcp[192.168.0.4:13405-192.168.1.3:23]

在实验中我们所采用的数据是 KDD CUP1999[7]，这些数据是源自 1998 DARPA 入侵检测评估程序，这些测试数据是比较权威的入侵检测领域的数据. 这些数据分为 4 大类: R2U(Remote to User)攻击、DOS(Denial of Service)攻击、PROBING 攻击和 U2R 攻击.

3 聚类分析

聚类分析[8][9]是数据挖掘中一种重要的探测数据工具，其核心是聚类. 聚类指的是将对象划分成簇，同一个簇的对象具有相似的性质，而不同簇的对象则有很大不同. 聚类的核心是将数据集划分为不同的种类，由此识别异常行为的模式. 对于挖掘对象而言，可以通过其属性、特征或者与其他对象的关系对该对象进行描述，挖掘对象间的相似性依赖于测量的特征和距离度量，距离度量的定义对聚类算法的结果在相当程度上具有一定的影响. 蜜罐日志数据事先无法分辨入侵行为数据和正常行为数据，所以首先需要利用无监督聚类算法则对数据进行分类，把各种类数据标记出来.

由于 KDD CUP 数据既包含离散型数据又包含连续型数据，因此应首先对离散型的数据进行连续预处理，以便适应聚类算法的要求. 以 protocol_type 属性为例: 通过实验收集的数据 protocol_type 只有三个值: icmp，tcp，udp 均为字符型，不符合聚类算法的要求，应对数据进行进一步处理. 我们可以把 icmp，tcp，udp 进行编号为 00，01，11，同时把 protocol_type 也进行拆分为 protocol_type1 和 protocol_type2 两项，当 protocol_type1=icmp，对于的 protocol_type1=0，依此类推可将 protocol_type 属性由离散型属性转换为连续型属性. 详细的属性如表 2 所示.

表 2 聚类算法中的实验数据包含的类型及数量

数据组	包含类型及数量
第一组 R2U(共 6524 条)	multihop(33)，guess_password(3918)，phf(7)，Ftpwrite(8)，imap(4)，xsnoop(12)，sendmail(29)，warezmaster(2253)，xlock(15)，named(24)，normal(221)
第二组 DOS(共 6524 条)	Neptune(2577)，smurf(3739)，normal(208)
第三组 probing(共 6524 条)	nmap(389)，Ipsweep(1904)，Satan(2353)，portsweep(1668)，normal(210)
第四组 U2R(共 6524 条)	multihop(29)，guess_password(3913)，phf(3)，Ftpwrite(16)，imap(6)，xsnoop(20)，sendmail(27)，warezmaster(2243)，xlock(15)，named(26)，normal(226)

对于聚类算法中的实验数据包含的类型及数量分别有不同的特征属性对应值，具体如表 3 所示.

表 3 特征属性对应值

属性名称	对应值
protocol_type	将其映射到整数，对应为 1-4: 1(tcp); 2(udp); 3(icmp); 4(other)
Sevice	将其映射到整数，对应为 5-22: 5(ecr_i); 6(eco-i); 7(domain-u); 8(finger); 9(ftp-data); 10(ftp); 11(http); 12(hostnames); 13(imap4); 14(login); 15(private); 16(systat); 17(telnet); 18(time); 19(uucp); 20(netstat); 21(smtp); 22(others)
Flag	在此我们仅选择了两个变量: “REJ”和“SF”，将其映射到整数 23-25: 23(REJ); 24(SF); 25(others)
Src_bytes	将其映射到整数，对应为 26-32: 26(0); 27(1-50); 28(51-200); 29(201-500); 30(501-1000); 31(1001-3000); 32(>3000)，括号内数值单位 bytes
Dst-bytes	将其映射到整数，对应为 32-38，方法与 src-bytes 相似.
Land	将其映射到整数，对应为 39-40: 39(0); 40(1)

对于距离度量来讲，通过一个实例来说明. 在某一日志信息中，设 A 作为其网络协议属性，B 为源 IP 地址，C 为目的 IP 地址，t 为目的端口，定义对象间的运算如下:

若 A1=A2，则 A1-A2=0; 否则其差为 2;

若 B1=B2，则 B1-B2=0; 则 B1、B2 同属于一个网段，则其差为 1; 否则为 2;

若 C1=C2，则 C1-C2=0; 则 C1、C2 同属于一个网段，则其差为 1; 否则为 2;

若 t1=t2，则 t1-t2=0; 否则其差为 2.

两个对象间的距离应满足交换律，故对两个属性间的差值取绝对值. 网络连接属性的加权距离定义如下:

$$Distance(o_1-o_2)=k_1\times|A_1-A_2|+k_2\times|B_1-B_2|+k_3\times|C_1-C_2|+k_4\times|t_1-t_2|$$

其中，k1，k2，k3 和 k4 分别表示每对属性相似度的权重，其值通过对历史数据的拟合优化得到，在本例中 k1=1，k2=3，k3=3，k4=5.

4 数据挖掘算法构建及规则的提取

对数据进行在数据挖掘阶段，想要得到什么样的知识，就需要对已经准备好的数据源应用相应的挖掘算法. 蜜罐中记录的日志信息并不知道哪些是入侵数据，哪些是正常数据，本文采用无监督算法 K-means 算法[10]对数据进行分类，在把各类标记出来. 用 K-means 算法对蜜罐记录的数据进行分类建立在两个假设基础之上: 第一种假设是蜜罐记录的入侵行为与正常行为有较大的差异，可以通过相似度(属性)将其区分开来，第二种假设是正常行为的访问记录远远大于入侵行为的访问记录[11].

K-means 算法[10]是聚类分析的非监督学习算法之一，其基本策略为在数据中为每个簇随意选择找一个对象作为代表，从而在 n 个对象中确定 k 个簇. 这些代表被称为中心点(medoid)，其他对象为非中心点(non-medoid). 计算全体非中心点到每个中心点的距离，并把所有非中心点划分到和它距离最小最近的中心点所在的簇. 如果距离结果可以被改善，就不断地用非中心点代替中心点. 聚类的质量通过代价函数来评价，该代价函数反映了对象和它所属簇的代表之间的平均相异性:

(1)任选 k 个代表.

(2)计算每对代表对象 Oi 和非代表 Oh 间的代价之和 TCih.

(3)选择满足 min(Oi，Oh ，TCih)的 Oi，Oh 对，如果最小的 TCih 为负，用 Oh 代替 Oi，退回第 2 步.

(4)否则，为每个代表对象寻找最相似的代表.

算法包括以下几个步骤:

(1)任意选取 k 个对象作为初始中心点(代表对象)

(2)重复以下过程:

把剩余对象分配到距离最近中心点所在簇; 随机选择一个非中心点对象 O; 计算随机使用 O 交换 Oj 的总代价 S; 若 S<0，则用 O 交换 Oj，形成新的 k 个中心点的几何; 直到无变化发生

(3)结束

聚类生成之后进行标记分类. 在蜜罐收集的日志信息数据中，入侵行为远远大于正常行为，因此可以利用所有的类包含数据对象的个数排序. 假定一个比例 N，将包含数据量小于 N 的类标记为正常类，其余的标记为异常类. 标记的算法如下:

假设 Ci，i=1,2,...,num.cluster 为已生成的聚类，N 为 0~1 之间的一个常数;

步骤 1: Sort(Ci)，按照 Ci 中包含数据对象多少从大到小对 Ci 排序;

步骤 2: j=1; k=N* num.cluster;

步骤 3: repeat;

步骤 4: if j<k，then Ci 标记为正常类; else 将 Ci 标记为入侵类;

步骤 5: j++;

步骤 6: until j> num.cluster

对数据进行分类标记之后，对所有入侵行为按行为模式插入到规则链表，形成规则集，同时日志及管理控制服务器将网络上捕获的数据包与已有数据规则集进行匹配比较，删除和已有规则相同的，余下的便是新规则，并用其更新入侵规则库[12]. 通过应用该算法将收集数据分为三类: 正常行为的集合，入侵行为的集合和非正常行为的集合.

5 数据挖掘

用数据挖掘算法对已经分好类的数据进行有规则的挖掘，用相应的入侵提取规则对测试数据进行检测，来源于某高校校园虚拟 Honeynet 中日志服务器的数据经过挖掘和分析后，其实验结果数据如表 4 所示.

表 4 实验结果数据

类型	总数据量	正常数据	入侵数据
第一组 R2U	54000	46308	7692
第二组 DOS	54000	46308	7692
第三组 probing	54000	46308	7692
第四组 U2R	4320	4248	252

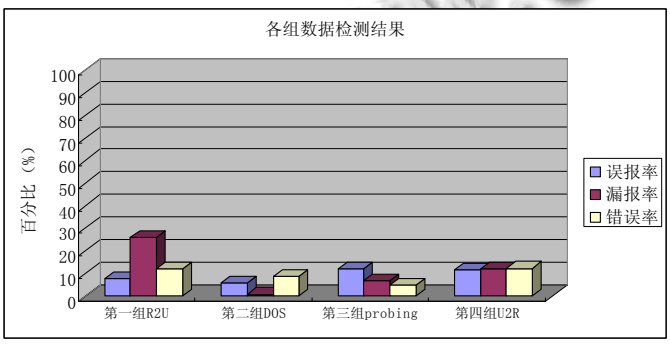

图 1 新入侵规则对各组数据的检测结果

在图 1 中，误报率=在正常数据中被错误认为入侵数据/正常数据记录总数，漏报率=入侵数据中被认为正常数据的数据量/入侵数据总量，错误率=(正常数据中的入侵数据+入侵数据中的正常数据)/全部测试数据.

从上面信息我们可以看出，由于 R2U 类数据中有较多数据是伪装合法的用户进行攻击，其各项特征与正常数据较为相似，造成算法很难将其区分，因此第一组 R2U 类入侵检测数据的检测不太理想，误报率、漏报率和错误率均高; 最后一组 U2R 入侵数据，可能由于入侵数据较少，对其的检测效果相对于 DOS 和 proing 类较差; 但该算法对 DOS 和 probing 类数据进行了有效的提取和检测，这表明使用聚类算法能够较好的把 DOS 和 probing 类入侵数据从正常数据中区分开来.

6 结 论

对于 Honeynet 中的日志信息数据进行挖掘是一个人机交互的过程，若 Honeynet 管理人员对得到的结果不满意，仍可重新定义相关的阈值，进行再次挖掘.

由于网络中的各个元素在不断变化，网络结构也不在断变化，黑客攻击的方法、模式、策略等也会相应地发生变化，Honeynet 收集到的有意义的规则则可能变得不再可用. 通过数据挖掘得到的多条规则可以更好的完善和补充规则库，一方面它可以如实地反映网络的变化而带来的攻击模式之间关系的变化，因此可将数据挖掘得到的结果作为预测攻击行为的依据; 另一方面它也相应地减轻了 Honeynet 管理人员的工作量，这也正是数据挖掘在 Honeynet 数据分析中的意义之所在.

参考文献

1 Cohen F. Deception Toolkit. http://www.all.net/dtk, 1998.

2 Olivier T, Marc D. A framework for attack patterns' discovery in honeynet data. Digital Investigation, 2008, 5(1): 128–139.

3 Cai JY, Vinod Y, Chris A,et al. Honeynet games: A game theoretic approach to defending network monitors. Journal of Combinatorial Optimiztion, 2011, 22(3): 305–324.

4 林杰斌,刘明德,陈湘.数据挖掘与 OLAP 理论与实务(第 2 版).北京:清华大学出版社,2002:35–54.

5 王实,高文.数据挖掘中的聚类方法.计算机科学,2000,(4):75–79.

6 赵会锋,李丽娟.一种基于蜜网的网络安全联动模型.计算机系统应用,2011,20(11):128–130.

7 张新有,曾华燊,贾磊.入侵检测数据集 KDD CUP99 研究.计算机工程与设计,2010(12):4809–4812,4816

8 程继华,施鹏飞.快速多层关联规则的挖掘.计算机学报,1998,(11): 1037–1041.

9 Soman KP, Diwakar S, Ajay V. Insight into Data Mining: Theory and Practice. New Delhi. Prentice-Hall of India Private Limited. 2009: 304–307.

10 刘明吉,王秀峰等.数据挖掘中的数据预处理.计算机科学,2000,27(4):54–57.

11 翟光群,陈向东,胡贵江.蜜罐与入侵检测技术联动系统的研究与设计.计算机工程与设计, 2009,30(21): 4847–4851.

12 金涛.数据挖掘在蜜罐日志分析中的应用研究[学位论文].上海:上海交通大学,2010.12:24–30