入侵检测系统的决策树生成方法

朱平哲

(三门峡职业技术学院 信息传媒学院,河南 三门峡 472000)

摘 要: 对网络攻击的检测一直是入侵检测系统研究的重要课题,故提出了一种入侵检测系统的决策树生成方法。首先,使用 DARPA 98 林肯实验室评估数据集作为训练和测试数据集;其次,描述了从 DARPA 数据集学习生成决策树的整个过程;最后,实验验证了决策树作为入侵检测系统数据挖掘方法的有效性。

关键词: 决策树;入侵检测系统;数据集;检测率

中图分类号: TP311.1 **文献标志码:** A **文章编号:** 1674-330X(2019)02-0072-05

Decision tree generation method in intrusion detection system

ZHU Pingzhe

(*Informational Media Department*, *Sanmenxia Polytechnic College*, *Sanmenxia* 472000, *China*)

Abstract: Detecting the network attacks is always an important topic during the researches of the intrusion detection system (IDS). A novel decision tree generation method in IDS is proposed in this paper. Firstly, the DARPA 98 Lincoln Laboratory Evaluation Data Set (DARPA Set) is used as the training data set and the testing data set. Secondly, the total process to generate the decision tree learned from the DARPA Sets is described. Finally, the effectiveness of the decision tree as the data mining method during the IDS is verified by a series of simulation experiments.

Keywords: decision tree; intrusion detection system; data sets; detection rate

DOI:10.16203/j.cnki.41-1397/n.2019.02.017

为了有效维护网络的稳定性和安全性,入侵检测系统(intrusion detection system,IDS)已成为连接在 Internet 上的网络安全基础设施的重要部分。IDS 有助于对网络入侵者进行检测、识别和跟踪,特别是基于网络的入侵检测系统(network-based IDS,NIDS)可以分析进入网络并受到保护的网络流量,以便对攻击进行检测和进一步分类。

目前,网络入侵检测方法主要可分为误用检测和异常检测[1]。大多数误用检测属于基于攻击的签名,在该攻击被识别和分析后,必须由安全专家加以定义,在定义签名后,将所有进入网络的业务与该签名进行比对,从而判断该业务是否试图攻击。因此,基于攻击特征的误用检测对于检测已知类型的攻击具有较好的性能。异常检测是对可疑业务与正常业务进行比较以应对攻击的一种方法,它的目标是在未知攻击成功之前对其加以阻止。为了检测异常流量,IDS 必须有自己的标准来感知流量是否属于攻击,用于异常检测的 IDS,首先要了解正常活动和异常活动的特征,然后检测偏离正常活动的流量[2]。为了对流量进行学习和分析,必须要发现流量数据集的分类规则[3]。因此,多种数据挖掘算法都可以作为异常检测的解决方案。

决策树是数据挖掘方法中的有效方法之一,它能够为 IDS 特别是异常检测提供合适的解决方案。因此,许多文献[4-6]均把决策树思想应用于入侵检测领域。为了将决策树作为异常检测的标准,需要训练数据集和评价数据集对决策树进行学习和评价。其中,DARPA 98 林肯实验室评估数据集(DARPA Set)就是 IDS 中一种被广泛使用的学习和测试数据集[7-8],它也被用于 KDD CUP 1999[9]。

尽管决策树和 DARPA 数据集在 IDS 中得到了广泛使用,但使用 DARPA 数据集生成决策树的过程一直

收稿日期: 2018-12-13
基金项目: 河南省教育厅科学技术研究重点项目(15B520026)
作者简介: 朱平哲(1982—),女,河南驻马店人,讲师,主要研究方向为智能信息处理。

以来并未得到清晰描述。为此,本方法将详细介绍如何利用 DARPA 数据集生成决策树,并对决策树作为异常检测方法的性能进行评价。

1　预备知识

1.1　决策树基本知识

(1) 决策树。决策树是最强大、最简单的数据挖掘方法之一。决策树由表示多个备选方案中每个选择的分支节点组成,每个叶子节点表示一类数据。决策树的一个简单示例如图 1 所示。

在图 1 中,诸如 T_1、T_2、T_3 和 T_4 的分支节点通过树中的测试向下过滤模式来向输入模式分配类号。例如,T_3 从 T_1 向下测试输入模式,并将类 3 分配给输入模式或传递给 T_4。最后,当输入模式到达叶子节点时,可将任何输入模式划分为类 1、2 或 3。因此,决策树对大数据集的数据分类是非常有价值的。

(2) 学习算法。决策树学习算法可以从学习的数据集中将特征定位于决策树中的合适位置,从而实现决策树的自动构建。目前有多重决策树学习算法,例如 ID3、C4.5 和 CART[10-12]。由于 ID3 算法采用了香农信息理论,概念清晰且容易实施,所以本方法采用 ID3 算法进行决策树的构建。

ID3 算法采用贪婪概念对决策树中的特征进行定位,即根据特征和类之间的相关性从学习数据集中选择特征。在信息论中采用熵的概念,如式(1)所示。式(2)中,信息增益用于衡量信息熵的预期减少:

$$ES = \sum_{i=1}^{S_{\max}} -p_i \log_2(p_i)\,, \tag{1}$$

式中: p_i 是数据集采用目标特征第 j 个数值时实例的比例。

$$Gain(S,A) = E_s - \sum_{v \in A} \frac{|S_v|}{|S|} E_{S_v}, \tag{2}$$

式中: v 是特征 A 的值; $|S_v|$ 是 S 实例的子集,A 取值为 v; $|S|$ 是实例的数量。

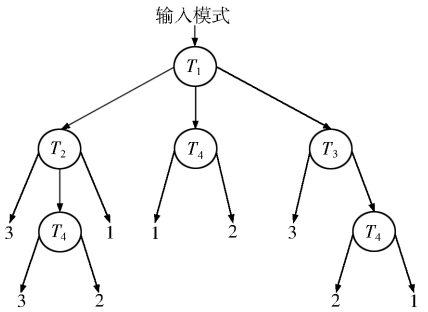

图 1　决策树简单示例

Fig.1　Simple example of decision tree

1.2　DARPA 98 林肯实验室评估数据集

DARPA 数据集是由麻省理工学院林肯实验室的信息系统技术组定义的,由美国国防高级研究计划署和空军研究实验室赞助。它是用于学习和测试的数据集,在各种 IDS 研究中有着广泛应用,对于比较评价结果非常有用。

用于学习的 DARPA 数据集包括 7 周的数据,每周有 5 天,每天有 BSM 审计数据和 TCP 转储数据。它还提供了 TCP 转储列表文件,该文件标记每个流,无论该流是否受到攻击。DARPA 集中的所有攻击可分为 4 类,具体类型如下:

(1) Denial of Service(DoS): 拒绝服务攻击,例如 TCP 洪水攻击(SYN flood);

(2) Remote to Local(R2L): 尝试从远程计算机进行未经授权访问的攻击,例如猜密码攻击;

(3) User to Root(U2R): 试图对本地超级用户(root)特权进行未经授权访问的攻击,例如各种缓冲区溢出攻击;

(4) 扫描: 非法监视和获取信息的攻击,例如端口扫描攻击。

本方法只考虑 TCP 转储数据和 TCP 转储列表文件,并使用 DARPA 数据集来学习和评估决策树。

2　决策树生成过程

决策树的生成过程如图 2 所示。图 2 中,预处理过程接收 DARPA 训练集作为输入,并为 ID3 算法的输入数据适当生成每个特性的值,ID3 算法负责选择特征并生成决策树。

2.1 DARPA 数据集划分

在 DARPA 训练集中,攻击主要分为 4 类。因此,在整个 DARPA 训练集中提取每个攻击的 TCP 转储数据,并为每个攻击类生成 4 个决策树。

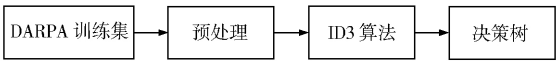

图 2 决策树生成过程

Fig.2 Process to generate the decision tree

表 1 给出了 TCP 转储列表文件的一部分。其中,每个条目均由流标识符号、日期、到达流的第一个包的到达时间、持续时间、服务名称、源端口号、目的地端口号、源 IP 地址、目的地 IP 地址、攻击得分和攻击名称组成。基于这个文件,能够识别哪个流是攻击,并能够利用 TCP 转储列表文件中的信息从 TCP 转储数据中提取数据。例如,使用表 1 中的信息从 TCP 转储文件中提取 smurf 攻击的命令如下:

①tcp dump-r input.dump src host 202.77.162.40 and dst host 172.16.114.50-w smurf1.dump;

②tcp dump-r input.dump src host 202.77.162.178 and dst host 172.16.114.50-w smurf2.dump。

通过这个过程,可以从 DARPA 训练集的 TCP 转储文件中提取各种攻击。

表 1 TCP 转储列表文件的采样

Tab.1 Sample of TCP dump list file

流标识符号	日期	到达流的第一个包的到达时间	持续时间	服务名称	源端口号	目的端口号	源 IP 地址	目的 IP 地址	攻击得分	攻击名称
1687	1998-07-02	08:16:57	0:00:01	snmp/u	161	1411	192.168.1.1	194.27.251.21	0	-
1688	1998-07-02	08:16:57	0:00:01	snmp/u	1411	161	192.27.251.21	192.168.1.1	0	-
1689	1998-07-02	08:16:59	0:05:28	ecr/i:r1215	-	-	202.77.162.40	172.16.114.50	1	smurf
1690	1998-07-02	08:16:59	0:05:28	ecr/i:r1206	-	-	202.77.162.178	172.16.114.50	1	smurf

2.2 预处理

假设目前已拥有各种攻击的 TCP 转储文件,但这些文件还未做好处理成为 ID3 算法的输入。由于无法处理 ID3 算法的连续值,故必须对 TCP 转储文件进行预处理,使其成为 ID3 算法的合适数据。不仅如此,预处理还有助于汇总来自 TCP 转储文件的信息。这里并不使用 TCP 转储文件中包含的所有信息,而是制造原始数据包数据使信息更有意义,即所谓的“选择特性”,所选择的特征必须很好地识别分组的特征。本方法选取的特征是 Snort [13] 主要用于检测攻击的属性。这里选择 5 元组特征、IP TOS、IP 长度、IP 分段、IP TTL、UDP 长度、TCP 标志、TCP 窗口大小、TCP 紧急指针、ICMP 类型、ICMP 代码、每秒数据包(packets per second,PPS)和每秒比特数(bits per second,BPS)。结合这些特征对原始分组数据进行了综合,即编码,这些编码规则必须能很好地识别分组报头字段的特征,如表 2 所示。

2.3 学习数据

学习数据是 ID3 算法的输入,ID3 算法通过前述过程将数据与编码后数据混合。学习数据必须包含正类和负类的数据。正类的数据是目标攻击类的数据,负类的数据是除目标数据之外的任何数据,本方法用非攻击数据的正常数据和不包括在目标攻击类中的攻击数据组成了负类的数据。例如,组成用于 DoS 攻击的学习数据包含 DoS 攻击的每个编码数据,例如 ping of death(POD)、smurf、land、teardrop 等。DoS 攻击的学习数据还包含作为无攻击数据的编码正常数据和可能属于 R2L、U2R 和 Scan 类的编码非 DoS 攻击数据。表 3给出了学习数据的一小部分。从表 3 中不难发现,学习数据中的每个条目必须都具有每个特性的值,并且指示条目被分为正类和负类。如果是针对每个攻击类的学习数据,则可以通过将学习输入 ID3 算法来生成每个攻击类的决策树。

2.4 生成树

限于篇幅原因,无法对每个决策树进行图形定位。因此,仅生成 U2R 和 R2L 攻击的决策树。

表 2　编码规则示例

Tab.2　Example of encoding rule

报头字段	编码规则	编码代码
IP TOS	*TOSfield* = 0	IP TOS = 1
	TOSfield > 0	IP TOS = 2
IP TTL	*TTLfield* < 64	IP TTL = 1
	64 < *TTLfield* ≥ 128	IP TTL = 2
	128 < *TTL* ≥ 192	IP TTL = 3
	192 < *TTL* ≥ 255	IP TTL = 4
TCP PORT #	*port* = 80	PORT = 1
	port = 20	PORT = 2
	port = 21	PORT = 3
	port ≥ 49 151	PORT = 24
PPS	*pps* ≤ 20	PPS = 1
	pps < 100	PPS = 2
	pps < 300	PPS = 3
	pps ≥ 300	PPS = 4

表 3　学习数据示例

Tab.3　Example of learning data

数据条目	分类
1 1 2 3 1 1 3 23 5 0 1 1 0 0 0 0 0 1 1	yes
1 1 2 3 1 1 3 23 5 0 1 1 0 0 0 0 0 2 1	yes
1 1 2 2 3 1 6 0 0 0 0 0 0 0 0 0 0 1 1	yes
1 1 2 3 1 1 3 23 5 0 1 1 0 0 0 0 0 2 2	yes
1 1 2 3 1 1 3 5 23 0 1 1 0 0 0 0 0 2 2	no
1 1 1 1 1 2 3 5 23 0 1 1 0 0 0 0 0 1 1	no
1 1 1 3 1 1 3 23 22 0 1 1 0 0 0 0 0 2 1	no
1 1 1 3 1 1 3 22 23 0 1 1 0 0 0 0 0 2 1	no
1 1 2 3 1 1 3 23 5 0 3 1 0 0 0 0 0 1 1	no

3　实验仿真

采用 ID3 算法和 DARPA 训练集生成每个攻击类的决策树，对决策树加以评估，测试数据集是 DARPA 测试集。DARPA 测试集在形式上与 DARPA 训练集相同，只是 DARPA 测试集包含比 DARPA 训练集更多的攻击。因此，决策树可以用新的攻击类型进行测试，这意味着决策树可以作为异常检测方法进行测试。

3.1　DoS 攻击检测率

在以往的 DoS 攻击中，back 攻击、land 攻击和 neptune 攻击的检测率均超过 90%，如图 3(a) 所示。然而，其他攻击的检测率低于 50%。

pod 攻击、smurf 攻击和 smurfttl 攻击主要使用 ICMP。与使用 UDP 或 TCP 的 DoS 攻击相比，使用 ICMP 的数据包属于 DARPA 训练集中较小的部分。因此，由于生成决策树的 ID3 采用了信息熵的概念，所以使用 ICMP 的分组信息不能充分影响决策树。

类似地，neptunettl 攻击的信息也不足以在决策树中加以显示，因为 neptunettl 攻击中的大多数数据包都指向 telnet 端口，但指向 telnet 端口的大量数据也包含在否定类的数据中。

图 3(b) 给出了新型 DoS 攻击的检测率。不难看出，尽管出现了新的攻击类型，但是 apache-2 攻击的检测率为 100%。这是因为用于 apache-2 攻击的编码数据模式与旧的 DoS 攻击相似。然而，由于编码数据的模式与旧的 DoS 攻击模式有很大不同，所以很少能检测到其他攻击，如邮箱攻击、进程表攻击和 UDP 风暴。

3.2　R2L 攻击检测率

R2L 攻击检测率见图 4。如图 4 所示，imap 攻击和 phf 攻击的检测率为 100%。ftp-write 攻击的检测率也很高，而 dict 攻击和 guest 攻击的检测率接近 50%。dict 攻击和 guest 攻击的编码模式很难与正常数据相对照。因此，许多特征如 TCP 端口号、IP 长度、IP TTL 和 TCP 窗口大小均被用于对这些攻击进行分类。然而，与 imap 和 phf 攻击相比，这些攻击的检测率较低。为解决该问题，定义和采用了比目前更多的特性，但这也导致生成了复杂程度更高、性能更低的决策树。

除蠕虫攻击和 xsnoop 攻击外，针对新型 R2L 攻击的检测性能不佳。蠕虫攻击和 xsnoop 攻击都具有前面定义的特征，而其他攻击则没有。低检测率的攻击具有与 U2R 攻击或扫描攻击类似的编码模式，而 R2L 攻

击则没有。为了克服这个问题,需要更多能够检查数据包内容的特性。

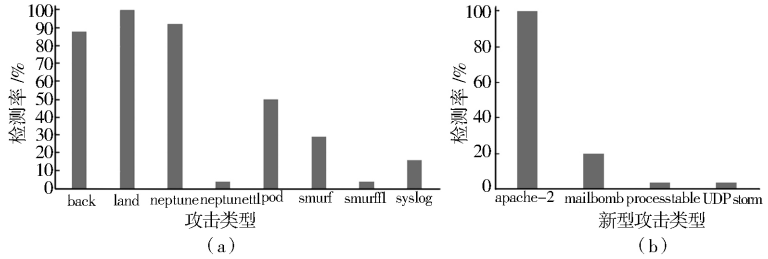

图 3 DoS 攻击检测率

Fig.3 Detection rate for DoS attack

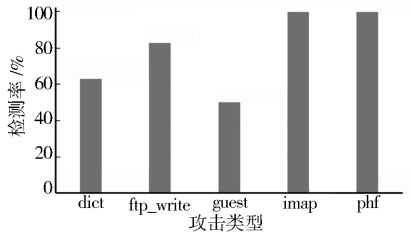

图 4 R2L 攻击检测率

Fig.4 Detection rate for R2L attack

3.3 Scan 攻击检测率

对于 Scan 攻击,新旧攻击类型的检测率均较高。扫描攻击的数据量大于其他攻击类中包含的攻击,故扫描攻击类的模式比其他攻击类多样化,这意味着可以提供大量的数据模式作为学习数据集。因此,扫描攻击可以获得较高的检测率,如图 5 所示。

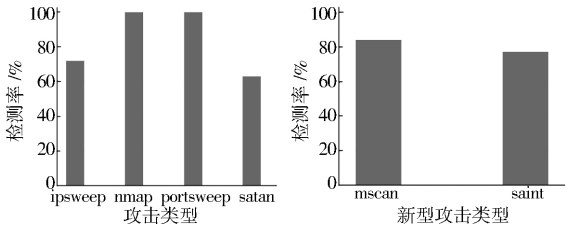

图 5 Scan 攻击检测率

Fig.5 Detection rate for Scan attack

4 结语

提出了用于检测 DoS 攻击、R2L 攻击、U2R 攻击和扫描攻击的决策树生成方法。该方法采用 ID3 算法作为学习算法自动生成决策树,训练数据采用 DARPA 数据集进行。这些方法虽然在 NIDS 的异常检测中得到了广泛应用,但对于决策树的整个生成过程缺乏描述。本方法描述了逐步生成决策树的过程,通过 DARPA 数据集测试数据并对决策树进行了评估。因此,该模型在检测新的攻击类型(异常检测)方面有了较大的改进。在下一步的工作中,可以定义能够表征分组内容和报头信息的更详细的特征,以期进一步提升该方法的性能。

参考文献:

[1] 胡博.面向异常检测的双重否定黑洞覆盖算法[J].南京理工大学学报(自然科学版),2018,42(5):604-608.

[2] KEMMERER R A,VIGNA C.Intrusion detection: A brief history and overview[J].Computer,2002,35(4):27-30.

[3] QUINLIN J R.Decision trees and decision making[J].IEEE Transactions on System,Man and Cybernetics,1990,20(2):339-346.

[4] ABBES T,BOUHOULA A,RUSINOWITCH M.Protocol analysis in intrusion detection using decision tree[C]//International Conference on Information Technology: Coding & Computing,Las Vegas,NV,USA,IEEE,April,2004: 404-408.

[5] KRUEGEL C,TOTH T.Using decision trees to improve signature-based intrusion detection[M].Berlin: Springer,2003.

[6] GARCIA V H,MONROY R,QUINTANA M.Web attack detection using ID3[M].Boston: Springer,2006.

[7] 翟继强,马文亭,肖亚军.Aprior-KNN 算法的警报过滤机制的入侵检测系统[J].小型微型计算机系统,2018,39(12):2632-2635.

[8] 徐慧,方策,刘翔,等.改进的飞蛾扑火优化算法在网络入侵检测系统中的应用[J].计算机应用,2018,38(11): 3231-3235,3240.

[9] ACM KDD Cup[EB/OL].(1999-01-13)[2018-12-30].http://www.sigkdd.org/kddcup/index.php.

[10] JEARANAITANAKIJ K.Classifying continuous data set by ID3 algorithm[C]//5th International Conference on Information Communications & Signal Processing, Bangkok,Thailand,IEEE,December,2005: 1048-1051.

[11] RUGGIERI S.Efficient C4.5[J].IEEE Transactions on Knowledge and Data Engineering,2002,14(2):438-444.

[12] SAFAVIAN S R,LANDGREBE D.A survey of decision tree classifier methodology[J].IEEE Transactions on Systems,Man and Cybernetics,1991,21(3):660-674.

[13] Web page of Snort[EB/OL].(2001-01-13)[2018-12-30] http://www.snort.org.