DOI:10.16357/j.cnki.issn1000-5862.2010.03.023
第34卷第3期 江西师范大学学报(自然科学版) Vol.34 No.3
2010年5月 JOURNAL OF JIANGXI NORMAL UNIVERSITY(NATURAL SCIENCE) May 2010

文章编号:1000-5862(2010)02-0302-06

基于决策树的协同网络入侵检测

蒲元芳, 张 巍, 滕少华, 杜红乐
(广东工业大学 计算机学院,广东 广州 510006)

摘要:由于不同网络协议有不同的属性值,不同的数据集可被用来检测网络入侵.该文提出了一种基于决策树的协同网络入侵检测模型,该模型是由多个代理组成,每个代理针对不同的网络数据协议类型(TCP/UDP/ICMP)分别履行检测,且它们又通过协同构成一个整体检测体系.最后用KDD CUP 99数据进行实验,验证了该方法检测入侵行为的有效性.

关键词:决策树;协同;网络入侵检测;协议类型
中图分类号:TP 183 文献标识码:A

0 引言

入侵检测系统(Intrusion Detection System,IDS)作为一种动态网络安全技术是当前网络安全研究的重点.根据检测方法,入侵检测分为2大类型:①误用检测:运用已知攻击方法,按照已定义好的入侵模式,通过判断这些入侵模式是否出现来检测;②异常检测:将偏离正常行为模式视为入侵嫌疑,即任何不符合以往活动规律的行为都将被视为入侵嫌疑[1].随着高速网络的出现,网络数据的流量不断增加,攻击方法的变化和不断更新又增加了检测攻击的难度.面对海量数据和复杂的攻击行为,需要更好的数据挖掘技术和检测方法才能应对入侵行为.决策树方法作为数据挖掘中分类技术的一种,具有分类简单、快速,且有较高准确率的特点[2-3],而入侵检测问题本质上是一个分类问题,所以将决策树分类方法应用于入侵检测得到了广泛的研究.

国外方面,Makkithaya等[4]提出了基于模糊聚类的决策树模型;Stein等[5]提出基于遗传算法的属性选择决策树模型来进行入侵检测.在国内该领域研究也发展迅速,宋明秋等[6]将决策树与协议分析结合起来研究入侵检测;赵晓峰等[7]将权值引入到随机决策树的构建中,给出了一种加权多随机决策树方法.以上方法通过实验研究分析,取得了较好的检测效果,证实了其可行性.本文提出了一种基于决策树的协同入侵检测模型,针对不同协议类型分别构建决策树进行入侵检测,加快检测入侵行为的速度以及提高分类的准确率,并对KDD CUP 99数据集进行实验,验证了该方法的有效性.

1 决策树算法

1.1 决策树简介

决策树是一个类似于流程图的树结构,每个内部节点表示一个属性上的测试,每个分支代表一个测试的输出,而每个叶子节点存放一个类标号,树的顶层节点是根节点.决策树的构造从代表全部训练样本的根节点开始,为每个内部节点选择一个分裂属性,并根据该属性的取值将样本划分为若干分支,直到叶节点将样本划分为某一类.

最早出现的决策树算法是1966年由Hunt等人提出的CLS算法,其主要思想是从一棵空的决策树开始,

收稿日期:2009-12-09
基金项目:广东省自然科学基金(06021484,9151009001000007,9451009001002777)和广东省科技计划(2008A060201011)资助项目.
作者简介:蒲元芳(1984-),女,湖北十堰人,硕士研究生,主要从事数据挖掘、机器学习研究.

通过添加结点逐步求精, 直到产生一棵能正确分类训练实例的决策树为止. 由于 CLS 算法在构造决策树的过程中没有给出选择测试属性的具体标准, 因此在该算法的基础上, Quinlan 提出了在国际上具有影响力的以信息增益作为属性选择度量的 ID3 算法[8] 和改进的 C4.5 算法[9] . 大约同时期, 几位统计学家 L. Breiman 等提出了著名的 CART 算法, 以 Gini 指标作为属性选择度量. 此外, 常见的决策树算法还有 CHAID、SLIQ 和 SPRINT 等.

1.2　C4.5 算法

这里主要介绍在国际上具有影响力的 C4.5 算法, C4.5 算法是在 ID3 算法的基础上改进的. ID3 使用信息增益作为属性选择度量. 设 D 为类标记元组的训练集, 类标记有 m 个取值表示 m 个类 $C_i(i = 1, \cdots, m)$, p_i 是 D 中任意元组属于类 C_i 的概率. 对 D 中元组分类所需的期望信息由 $Info(D) = -\sum_{i=1}^{m} p_i \log_2 p_i$ 给出, $Info(D)$ 是识别 D 中元组的类标号所需要的平均信息量, 又称为 D 的熵.

根据属性 A 将 D 划分为 v 个子集 $\{D_1, D_2, \cdots, D_v\}$, 理想情况是该划分产生元组的准确分类, 这对数据集提出了很高的要求, 即每个数据集都是纯的, 然而一般来说数据不纯. 因而, 为了得到更准确的分类结果, 本文引用信息增益算法, 按 A 划分对 D 的元组分类所需要的期望信息计算公式为 $Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$.

信息增益定义为原来需要的信息量与新的需求之差, 其计算公式为 $Gain(A) = Info(D) - Info_A(D)$. 选择具有最高信息增益 $Gain(A)$ 的属性 A 作为分裂属性等价于按能做"最佳分类" 的属性 A 划分, 使得完成元组分类还需要的信息量最少.

由于信息增益度量偏向具有许多输出的测试, 即倾向于选择具有大量值的属性, 为了克服这种偏倚, Quinlan 随后提出以信息增益率为属性选择度量的 C4.5 算法. C4.5 对 ID3 做了改进, 以信息增益率作为属性选择度量. 首先使用分裂信息将信息增益规范化, 分裂信息定义由公式 $SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2(\frac{|D_j|}{|D|})$ 给出, 该值代表通过将 D 划分成对应于属性 A 测试的 v 个输出, 产生 v 个划分的信息.

信息增益率是信息增益与分裂信息的比例, 由 $GainRatio(A) = Gain(A)/SplitInfo(A)$ 给出, 选择具有最大增益率的属性作为分裂属性.

2　基于决策树的协同入侵检测模型

由于入侵检测的实质是根据提取到的网络数据信息将网络行为分类为正常行为和入侵行为[10-12], 而网络行为的识别离不开网络协议. 不同网络协议表现出不同的网络连接特征, 如面向连接的 TCP 协议, 面向无连接的 UDP 协议以及 ICMP 协议. 本文针对不同的协议构建决策树模型, 使之协同作业以减少检测时间并提高分类的准确率. 构建的模型如图 1 所示, 共分为 4 个模块: 数据采集, 数据预处理, 基于决策树的检测代理,

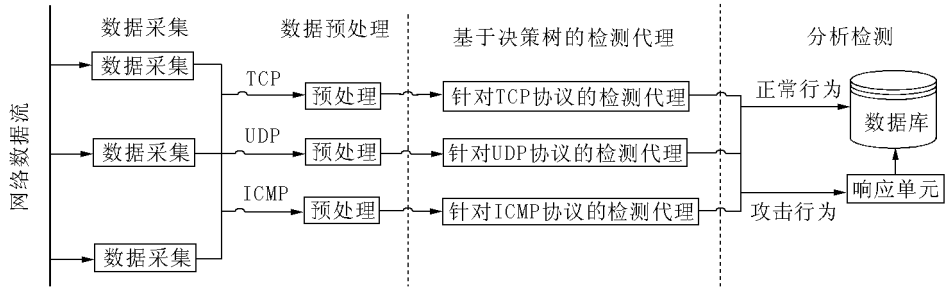

图 1　基于决策树的协同入侵检测模型

分析检测.

数据采集模块用来收集网络数据;数据预处理模块针对网络数据进行分流、预处理、属性选择及格式转换等工作;基于决策树的检测代理是针对不用协议类型的数据分别建立决策树模型并分类;分析检测模块则针对多个检测模型的分类结果,将正常数据存放到数据库中,如果发现攻击行为,响应单元则采取相应的响应措施处理入侵.

3 基于决策树的检测代理

针对3种协议类型的网络数据包,将网络数据流分成3种,分别为TCP数据流、UDP数据流和ICMP数据流,然后根据3种网络协议数据的特点分别构建基于决策树的检测代理来履行TCP、UDP和ICMP攻击检测.每个决策树检测代理的建立都经历了数据预处理、建立模型、模型评估及检测过程.以针对TCP协议数据的检测代理为例具体说明,针对UDP和ICMP协议的网络数据检测代理有类似的情况.针对TCP协议数据建立的基于决策树的检测代理如图2所示.

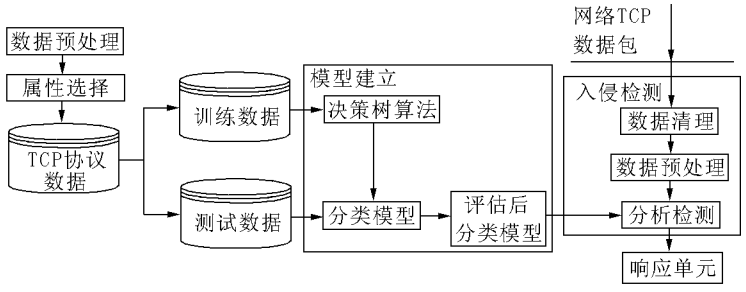

图2 针对TCP协议数据的决策树检测代理

3.1 数据采集和预处理

要检测来源于网络的入侵,首先需收集网络数据.由于现实中的数据一般是"脏"的、不完整的和不一致的,数据预处理技术可以改进数据质量,高质量的决策依赖于高质量的数据,因此当数据采集完成之后将对数据进行预处理,主要任务有数据清理、数据集成、数据变换、数据归约和数据离散化,如图3所示.

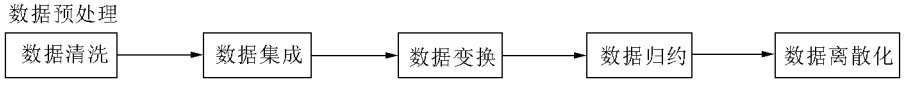

图3 数据预处理步骤

数据清理通过对收集的数据填写缺失值,光滑噪声数据、识别或删除信息量非常小的数据,解决数据的不一致性.数据集成合并不同数据采集点收集的数据,按TCP、UDP和ICMP协议分类存放.数据变换、归约和离散化则是为了将原始数据处理为各种形式的结构化且富含语义的记录,致使它们可以产生准确、有用且易于理解的结果.例如由tcpdump输出的原始数据是二进制编码,需要进行处理,将其转换成包含多个属性的形式,如包含源IP地址、目标IP地址、源端口、目标端口、协议类型等然后再进行下一步的操作.

经过预处理后的数据将进行属性选择,通过删除不相关或冗余的属性来减小待分析的数据集,如TCP、UDP、ICMP协议格式不同,UDP检测代理就不需要TCP格式的一些属性.由于在3种协议中分别有部分属性只有唯一值,即每条记录属性取值相同,故不能作为分裂属性.经排除,用41个属性中的35个属性进行实验来应对TCP攻击的检测,20个属性来应对UDP的攻击检测,16个属性来应对ICMP攻击检测.如果不分协议类型进行处理,那么在整个数据集中只有3条属性有唯一值,故排除后还有38个属性参与构建决策树.因此根据不同协议类型的数据包分别建立检测模型,可以大大减少待分析的数据集,有助于提高分类的准确率及减少检测模型的构建时间.

3.2 决策树模型的建立

针对不同协议的数据,分别建立决策树的检测代理检测入侵.为了建立有效的检测代理,避免出现过度

适应的情况, 将预处理后已知行为的数据集随机地划分成 2 个独立的集合: 训练集和测试集. 一般 2/3 的数据作为训练集, 其余 1/3 作为测试数据. 训练集用于建立模型, 测试集用于评估模型的准确率.

建立决策树模型时要选择合适的决策树算法, 目前流行的决策树算法有很多, 但 ID3 算法只能处理连续属性, C4.5 算法既能处理连续属性也可以处理离散属性, CART 算法主要强调树是二叉的. 针对网络数据的情况, 既有连续属性又有离散属性, 因此不能使用 ID3 算法, 而且分类的类别大于 2 种, 因此不采用 CART 算法. 而 C4.5 算法发展比较完善而且简单易懂, 因此本文选用 C4.5 算法来构建决策树. 图 4 为以 KDD CUP 99 数据集中隔 100 行抽取 1 条记录, 共 49 401 条记录中的 TCP 协议数据为实验对象, 利用数据挖掘工具 Rapid-Miner 建立的决策树.

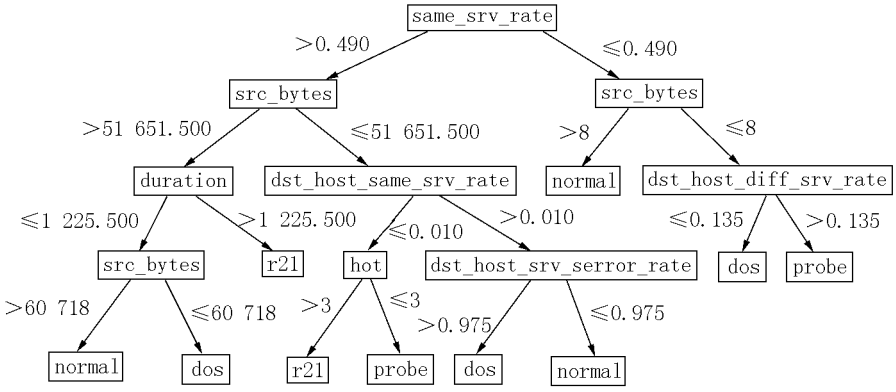

图 4　针对 TCP 协议的数据构建的决策树

3.3　模型评估

对于建立后的模型, 需要用测试集对模型评估, 评估是对分类准确性的一个评判, 以检测率为主要度量. 检测率是能够检测出的入侵行为的概率:

检测率= 被正确检测为入侵的记录数/实际测试入侵的记录数.

检测率过低会导致漏报率过高, 降低被保护系统的安全性, 因此只有具有较高检测率的模型才是有效的模型, 才能用于实际的网络数据入侵检测.

3.4　分析检测

由于决策树模型产生的结果用规则的形式表示, 规则存放在特征规则库中, 对于处理后的数据, 利用模式匹配, 把数据特征与规则库中的规则进行比较. 对于误用检测, 当发现数据特征与某个已知入侵模式匹配时, 则当前行为为已知攻击或入侵. 对于异常检测, 当数据特征偏离正常行为模式时, 则视为入侵嫌疑. 一旦发现攻击则发出相应的报警.

4　实验分析

4.1　数据来源

本实验采用的数据集来自 KDD CUP 99, 是美国 MIT Lincoln 实验室通过模拟一个典型的美国空军网站并将 9 个星期所收集的原始数据加工而成的 500 万条样本记录. 包含 TCP、UDP、ICMP 等 3 种协议, 每条记录包含 41 个属性和一个类别标记, 类别标记除了正常数据外还有 4 种攻击类型, 分别为 DoS(Denial of Service Attacks)、Probe、U2R (User to Root Attacks)、R2L (Remote to User Attacks). 从中隔 100 行抽取 1 条记录, 共 49 401 条记录作为实验对象.

4.2　数据分析

对抽取的 49 401 条记录, 按协议类型统计数据分布情况如表 1 所示. 由表 1 中的数据可以看出, DoS 攻击主要出现在 ICMP 协议中, 其次是 TCP 协议; Probe 攻击主要出现在 TCP 协议中, 其次是 ICMP 协议; 而 R2L 和 U2R 攻击只出现在 TCP 协议中. UDP 协议中出现的攻击行为比较少, 而 ICMP 协议中绝大部分情况都属于

攻击.

由于 U2R 攻击的数据量太少,仅占整个数据集的 0.001%,即数据集中任一记录属于 U2R 类的概率极低,在构建决策树时不利于计算信息增益,因此本文中不讨论 U2R 攻击的情况,只对其他 4 种类型进行分类.

表 1 数据类别分布情况统计

数据集	Normal	Dos	Probe	R2l	U2r	总计数目
全部抽取数据	9 721	39 144	415	117	4	49 401
TCP 协议数据	7 682	10 942	267	117	4	19 012
UDP 协议数据	1 924	99	19	0	0	2 042
ICMP 协议数据	115	28 103	129	0	0	28 347

4.3 实验结果及分析

首先用抽取的 49 401 条记录用 C4.5 算法构建决策树,然后针对每种协议的数据分别用 C4.5 算法构建决策树.当把整个数据集分别分为正常数据和入侵数据 2 种类型和分为 Normal、Dos、Probe、R2l 等 4 种类型时,分类的准确率及构建决策树的训练时间分别如表 2、表 3 所示,其中针对 TCP 协议的数据构建的决策树如图 4 所示.

表 2 将数据类型分为 2 种类型的结果

数据集	正常数据/%	入侵数据/%	构建决策树的时间/s
全部抽取数据	99.69	99.77	9
TCP 协议数据	99.05	99.67	5
UDP 协议数据	100	97.43	4
ICMP 协议数据	100	100	4

表 3 将数据类别分为 4 种类型的结果

数据集	Normal/%	Dos/%	Probe/%	R2L/%	构建决策树的时间/s
全部抽取数据	99.17	99.83	89.06	72.73	9
TCP 协议数据	99.71	99.81	94.80	83.3	4
UDP 协议数据	100	100	83.3		3
ICMP 协议数据	100	100	100		4

从表中数据可以看出,不管是将数据分为 2 种类型还是分为 4 种类型,根据不同协议对数据分类都有比不分协议类型对数据分类具有更高的准确率,而且在构建决策树的时间上分别构建模型进行协同检测显然比针对全部数据构建决策树用的时间要少,因此协同作业不仅提高了数据分类的准确率也加快了入侵行为的检测速度.

5 结束语

决策树技术由于分类快速且准确率高的特点被广泛应用于入侵检测,本文提出了一种基于决策树的协同入侵检测模型,根据协议类型分别建立分类检测代理以达到协同检测的目的,提高分类的准确率并加快检测速度.最后以 KDD CUP 99 数据集为实验对象,验证了本文方法的可行性.下一步的研究工作将是如何针对 U2R 攻击进行检测以更好的识别 U2R 的攻击.

参考文献:
[1] 滕少华. 基于对象监控的分布式协同入侵检测[D]. 广州: 广东工业大学, 2008.
[2] Tan P, Steinbach M, Kumar V. Introduction to data mining [M]. Beijiang: Post &Telecom Press, 2006.
[3] Han J, Kamber M. 数据挖掘概念与技术 [M]. 2 版. 北京: 机械工业出版社, 2007.
[4] Krishnamoorthi Makkithaya, Subba Reddy N V, Dinesh Acharya U. Improved c-fuzzy decision tree for intrusion detection [J]. Engineering and Technology, 2008, 32: 140-145.
[5] Gary Stein, Bing Chen, Annie S, et al. Decision tree classifier for network intrusion detection with GA-based feature selection [D]. Fkorida: University of Central Florida Orlando.
[6] 宋明秋, 傅韵, 邓贵仕. 基于决策树和协议分析的入侵检测研究[J]. 计算机应用研究, 2007, 24(12): 171-173.
[7] 赵晓峰, 叶震. 基于加权多随机决策树的入侵检测模型[J]. 计算机应用, 2007, 27(5): 1041-1043.
[8] Quinlan J. Induction of decision trees [J]. Machine Learning, 1986(1): 81-106.
[9] Quinla J. C4. 5 programs for machine learning [M]. New York: Morgan Kauffman, 1993: 44-78.
[10] 邱舟强, 滕少华, 李振坤, 等. 数据挖掘技术在网络入侵检测中的应用[J]. 江西师范大学学报: 自然科学版, 2006, 30(2): 13-15.
[11] 滕少华, 王琳. 径向基神经网络在入侵检测中的应用[J]. 江西师范大学学报: 自然科学版, 2007, 31(3): 297-301.
[12] 张正球, 姚志强, 颜西山, 等. 适应性免疫的轻量级网络入侵检测算法[J]. 江西师范大学学报: 自然科学版, 2008, 31(6): 724-728.

The Cooperative Network Intrusion Detection Based on Decision Tree

PU Yuan-fang, ZHANG Wei, TENG Shao-hua, DU Hong-le
(College of Computer, Guangdong University of Technology, Guangzhou Guangdong 510006, China)

Abstract: Because different network protocols have different attributes, different data could be used to detect network intrusion. A cooperative network intrusion detection model based on decision tree is be proposed. This model is coposed of multi-agents, each agent is constructed for different protocol which is used to detect the network and they construct a detection architecture by cooperation. At last the data sets of KDD CUP99 are used as the experiment data, and the results show the efficiency of our method.

Key words: decision tree; cooperative; network intrusion detection; protocol type

(责任编辑: 冉小晓)