XUFE

数据挖掘在网络安全中的应用

YXLUO8836@163.COM

YangXia LUO

交流内容如下

1. 网络安全研究内容
2. 入侵检测相关问题
3. 目前的研究工作

一、网络安全研究内容

1. **网络安全管理:**

分散存储，要集中管理、集中监控和处理，统一审计、分布式结构

2. **建立安全模型**

综合技术结合：数据加密、网络侦听与反侦听等

立体化IP网络安全的研究

3. **有效的安全评估**

主要对系统的**脆弱性研究**，构造一个基于脆弱性扫描的网络安全评估方法，和设计了一个网络安全评估系统。

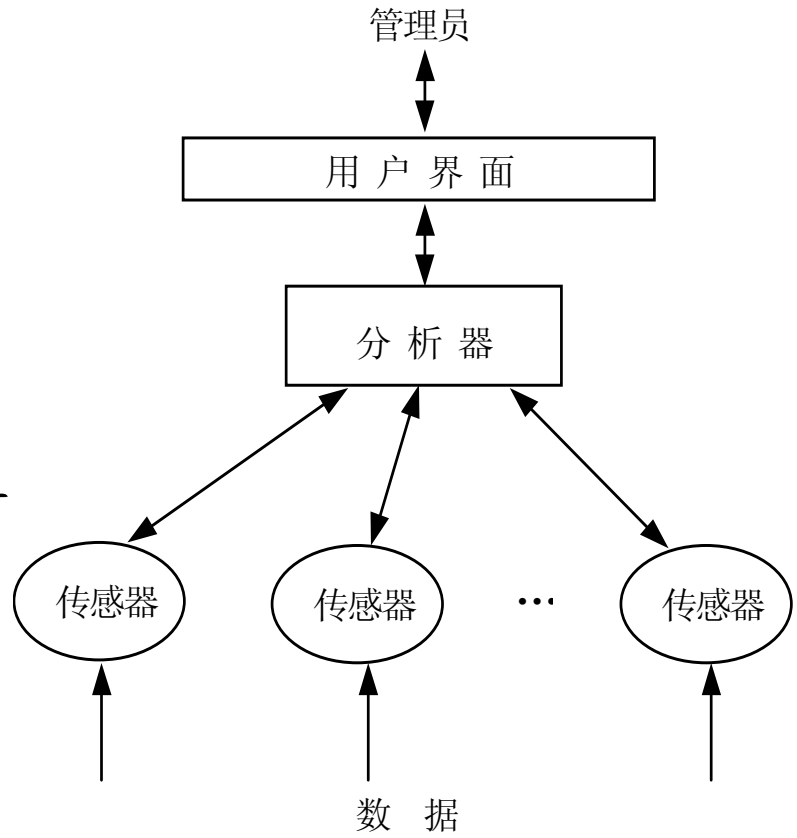
3. **智能入侵防御系统**

基于贝叶斯网络的，神经网络的，遗传免疫的，基于粗造集的，移动或智能Agent，基于虚拟网格的，分布式协同的

二、入侵检测相关问题

❖入侵检测工作

1. 收集信息（网络或主机）
2. 分析检测
3. 报警、结合防护系统驱逐
4. 增加防范系统知识库，增强防范能力



入侵检测技术

入侵检测系统

1. 特征检测(误用检测）——模式匹配

2. 异常检测--动态性，轮廓模型，阀值判定

特征检测分类：

- 状态转换法--状态迁移.
- 专家系统法：入侵特征表达为if-then结构的规则
- 模式匹配法：入侵特征编码成与审计记录相符合的模式,能够在审计记录中直接寻找到相匹配的已知入侵模式，

异常检测分类：

- 统计分析技术
- 机器学习
- 各种数据挖掘技术。

入侵检测特点

- **特征检测：** 依赖于模式库的完整性，对存在的模式检测精度高，对变体或新的攻击，漏报高。后验性，应用性强
- **异常检测**：具有动态性，具有智能性。先验性，研究性强

如：绿盟科技

入侵检测的数据

- **网络数据**

  收集网络中传输的报文，网络流量，协议端口，长度TCP和ICMP标志等。

- **主机数据**

  运行在网络中的关键主机上，收集操作系统数据，日志文件中的特征数据数据速率，连接等相关的变量

- **应用程序上的数据** 对数据中各标识项等。

三、目前研究方向和计划

- **实验目标：** 智能化入侵检测， 寻找一种或几种合理的检测算法，能够提高检测系统的智能化，先验性和检测异常的准确性。

依据数据挖掘的过程

1. 实验数据准备
2. 数据预处理
3. 对所选算法选择实验
4. 实验结果的比较分析
5. 模型转变成实际系统

1. 数据准备

实验数据 数据的采集可以通过一些抓包工具来获得，如

1. **Unix下的Tcpdump**
2. **Windows下的Libdump**
3. **专用的软件snort捕捉数据包**

生成连接记录作为数据源。

KDDCup99的网络入侵检测数据集

该数据集是从一个模拟的美国空军局域网上采集来的9个星期的网络连接数据，一共有490万条数据，

XUFE

数据集DKKCUP99

特征名	描　　述	类型
Duration	连接时间长度(单位：秒)	连续
protocol_type	协议类型，如 tcp，udp 等	离散
Service	在目标机的网络服务，如 http，telnet 等	离散
src_bytes	源地址到目标地址的数据流量	连续
dst_bytes	目标地址到源地址的数据流量	连续
flag	连接状态(正常或错误)	离散
land	1-数据连接源地址和目标地址为同一主机或端口；0-其他	离散
wrong_fragment	错误碎片的数目	连续
urgent	紧迫数据包的个数	连续

http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

输出结果：

与weka3.5通过jdbc方式相连：

0:0,tcp,http,SF,215,45076,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,0,0,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,normal.
1:0,tcp,http,SF,162,4528,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,1,1,1.00,0.00,1.00,0.00,0.00,0.00,0.00,0.00,normal.
2:0,tcp,http,SF,236,1228,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,2,2,1.00,0.00,0.50,0.00,0.00,0.00,0.00,0.00,normal.
3:0,tcp,http,SF,233,2032,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,3,3,1.00,0.00,0.33,0.00,0.00,0.00,0.00,0.00,normal.
4:0,tcp,http,SF,239,486,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,3,3,0.00,0.00,0.00,0.00,1.00,0.00,0.00,4,4,1.00,0.00,0.25,0.00,0.00,0.00,0.00,0.00,normal.
5:0,tcp,http,SF,238,1282,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,4,4,0.00,0.00,0.00,0.00,1.00,0.00,0.00,5,5,1.00,0.00,0.20,0.00,0.00,0.00,0.00,0.00,normal.
6:0,tcp,http,SF,235,1337,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,5,5,0.00,0.00,0.00,0.00,1.00,0.00,0.00,6,6,1.00,0.00,0.17,0.00,0.00,0.00,0.00,0.00,normal.
7:0,tcp,http,SF,234,1364,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,6,6,0.00,0.00,0.00,0.00,1.00,0.00,0.00,7,7,1.00,0.00,0.14,0.00,0.00,0.00,0.00,0.00,normal.
8:0,tcp,http,SF,239,1295,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,7,7,0.00,0.00,0.00,0.00,1.00,0.00,0.00,8,8,1.00,0.00,0.12,0.00,0.00,0.00,0.00,0.00,normal.
9:0,tcp,http,SF,181,5450,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00,0.00,0.00,0.00,1.00,0.00,0.00,9,9,1.00,0.00,0.11,0.00,0.00,0.00,0.00,0.00,normal.

标识类型	含义	具体分类标识
Normal	正常记录	normal
DOS	拒绝服务攻击	back、land、neptune、pod、smurf、teardrop
Probing	监视和其他探测活动	ipsweep、nmap、portsweep、satan
R2L	来自远程机器的非法访问	ftp_write、guess_passwd、imap、multihop、phf、spy、warezclient、warezmaster
U2R	普通用户对本地超级用户特权的非法访问	buffer_overflow、loadmodule、perl、rootkit

- 数据集中包含了1种正常的类型normal和22种训练攻击类型，另外有14种攻击仅出现在测试数据集中。
- 测试数据和训练数据有着不同的概率分布，测试数据包含了一些未出现在训练数据中的攻击类型，这使得入侵检测更具有现实性

其它数据集

- 其它实验dataset和UCI数据，如股票上，医院数据
  http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

australian	Statlog	classification	2	690		14
breast-cancer	UCI	classification	2	683		10
colon-cancer	[AU99a]	classification	2	62		2,000
covtype.binary	UCI	classification	2	581,012		54
diabetes	UCI	classification	2	768		8
duke breast-cancer	[MW01a]	classification	2	44		7,129
fourclass	[TKH96a]	classification	2	862		2
german.numer	Statlog	classification	2	1,000		24
heart	Statlog	classification	2	270		13
ijcnn1	[DP01a]	classification	2	49,990	91,701	22
ionosphere	UCI	classification	2	351		34
leukemia	[TG99a]	classification	2	38	34	7129
liver-disorders	UCI	classification	2	345		6

2、数据的预处理

只有就进行预处理，才可能挖掘出有用的模式，保证检验算法的效率和实验结果的正确。

- **数据清理**
  - 填写空缺的值，平滑噪声数据，识别删除独立点解决不一致问题
- **数据集成和变换**
  - 多数据库或文件的集成；规格化和聚集
- **数据归约**
  - 得到数据的压缩表示
- **数据离散化**
  - 针对连续数据值

在实验中对数据进行标准化和归一化后，再聚类。

3.数据挖掘中用于网络安全的算法

1. 基于贝叶斯算法
2. 基于支持向量机
3. 基于关联算法
4. 基于神经网络算法
5. 基于人工免疫算法

基于贝叶斯网络

贝叶斯网络:

是一种基于概率的不确定性推理方法，也是处理不确定性信息的主要工具。

研究方法:

把训练样本中所有的属性作为网络参数，通过训练样本，计算节点的概率。

- 提高系统入侵检测的效率
- 保证正确获得数据特征属性
- 减少了计算量，提高了分类的效率

基于SVM的入侵检测研究

❖ 支持向量机

一种分类和预测算法，解决异常检测的计算量较大的问题。

入侵检测问题本质上是一个分类问题，如果把多类分类问题转化为多个两类分类问题，我们就可以应用支持向量机解决攻击分类。

研究时可结合Hadamard矩阵，解决入侵检测的多类分类。

基于关联算法

多种关联规则算法

例如Apriori算法。其将关联规则的过程分为两个步骤:

1. 第一步通过迭代，检索事务数据库中的所有频繁项集，即支持度不低于用户设定的阀值的项集;
2. 第二步利用频繁项集构造出满足用户最小信任度的规则。

基于神经网络

基于无监督神经网络的入侵检测用于训练各种攻击的数据，这些数据并未注明或标识，而需要寻找和定义正常的聚类，在不具备任何先验知识的情况下发现新型攻击的能力。

方向:

- 提高对未知入侵的检测率。
- 利用自组织特征映射进行网络入侵检测，设计相应检测过程和算法。
- 自适应共振理论ART网络。
- 模糊自适应共振理论算法，提高入侵检测的可行性和有效性。

基于人工免疫的入侵检测

- 特点是利用免疫系统的原理、体系结构和相关算法来实现对入侵行为的检测。

研究的内容与目的 ：

1. 将免疫系统中的否定选择算法应用到入侵检测领域，提高系统的可扩展性和自适应性较差。
2. 对未知入侵行为进行检测。
3. 在入侵检测模型中建立免疫智能体的逻辑结构及其运行机制，目的实现主动防御机制。
4. 提高检测率高，误检率低。

其它研究方向

- 多层、立体化、协作式网络管理模型与技术研究
- 分布式拒绝服务攻击手段(DDOS)的研究
- 智能化、主动化入侵检测模型研究。

1. 基于可拓识别入侵检测模型，
2. 基于数据挖掘技术研究
3. 基于人工免疫算法研究
4. 基于网格技术的主动防御研究

感谢各位！