

# 12 先进的入侵检测技术

## 1 采用先进检测算法的必要性

## 2 神经网络与入侵检测技术★

## 3 数据挖掘与入侵检测技术★

## 4 数据融合与入侵检测技术

## 5 计算机免疫学与入侵技术

## 6 进化计算与入侵检测技术

### **1 采用先进检测算法的必要性**

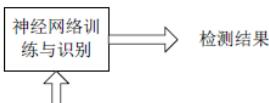
- 20世纪90年代以来，不少研究人员提出了不少新的检测算法，这些检测算法从不同角度看待入侵检测的基本问题，并利用许多人工智能或者机器学习的算法，试图解决传统入侵检测技术中存在的若干问题，如虚警、缺乏检测未知或变形攻击的能力、扩展性和自适应性等

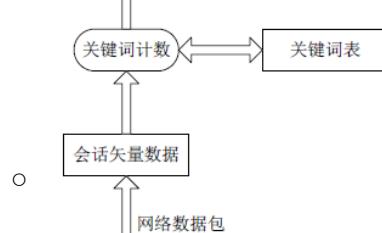
### **2 神经网络与入侵检测技术★**

- 人工神经网络是模拟人脑加工、存储和处理信息机制而提出的一种智能化信息处理技术
- 它是由大量简单的处理单元（神经元）进行高度互连而形成的复杂网络系统
- 从本质上讲，人工神经网络实现是一种从输入到输出的映射关系，其输出值由输入样本、神经元间的互连权值以及传递函数所决定
- 通过训练和学习来修改网络互连权值，神经网络就可以完成所需的输入—输出映射
- 神经网络技术应用于入侵检测领域有以下优势：
  - 神经网络具有概括和抽象能力，对不完整输入信息具有一定程度的容错处理能力
  - 神经网络具备高度的学习和自适应能力
  - 神经网络所独有的内在并行计算和存储能力
- 神经网络技术应用于入侵检测：
  - 采用递归型（Recurrent）BP网络
    - 最常用的神经网络就是BP网络，也叫多层前馈网络。BP是backpropagation的缩写，是反向传播的意思
    - 前馈是从网络结构上来说的，是前一层神经元单向馈入后一层神经元，而后面的神经元没有反馈到之前的神经元
    - BP网络是从网络的训练方法上来说的，是指该网络的训练算法是反向传播算法，即神经元的链接权重的训练是从最后一层（输出层）开始，然后反向依次更新前一层的链接权重
  - 基于一维SOM（组织特征映射）网络的异常检测算法对用户行为特征进行判断
  - 采用基于多层感知器（MLP）的异常检测模型
  - MIT提出了采用关键词和神经网络相结合的方法进行入侵检测并针对Telnet服务会话进行相关研究，采用的方法如下：
    - 选择一组关键词表
    - 在会话数据中对各个关键词进行计数并形成n维的输入特征矢量（n为关键词个数）
    - 采用MLP网络进行训练和识别
    - 对关键词的选取原则进行了一定的说明，并在达到80%检测概率的基础上将虚警概率降低到大约每天一次的水平
- 神经网络技术应用于入侵检测领域的缺陷：
  - 需要解决神经网络对大容量入侵行为类型的学习能力
  - 需要解决神经网络的解释能力不足
  - 执行速度

以Lippmann算法为例说明神经网络在入侵检测技术中的应用

- Lippmann算法如图所示：

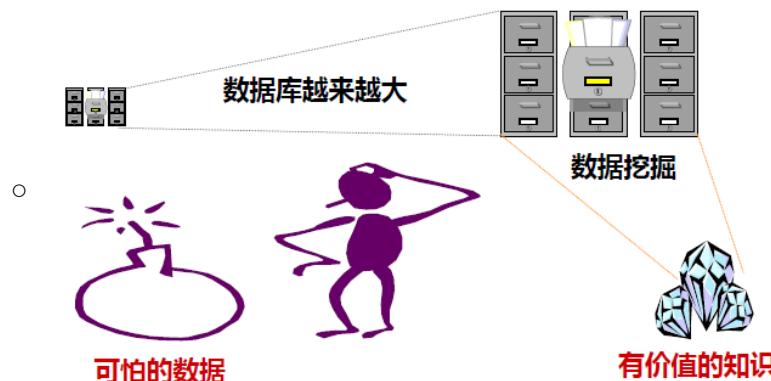




- Lippmann算法流程
  - 从所收集的网络数据包中构建网络会话数据矢量，即单个网络会话中双方实体之间所传输的所有数据负载内容。这里指的会话实体，可以理解成为网络编程中的套接字概念。不同的IP地址和端口的组合形成不同的会话实体
  - 根据选定的关键词表，在会话数据内容中搜索配匹的各个关键字，并形成各自的统计计数值
  - 将各个关键字计数值并行排列起来形成输入特征矢量
  - 在输入到神经网络进行训练之前，还必须进行预处理工作
  - 预处理工作完成后，输入特征矢量被送到神经网络模型中进行训练
- Lippmann对关键词的选择问题及其对检测性能的影响进行了研究
  - 选取适当类型的关键词
  - 加上简单的神经网络训练能力
- 能够较大地提高入侵检测的性能
- Lippmann的工作指出，该种检测方法对于训练样本集中未出现的新攻击类型，同样获得了较高的检测性能，体现出某种程度的攻击概括归纳能力

### 3 数据挖掘与入侵检测技术★

- 数据挖掘 (DataMining) 是所谓“数据库知识发现” (KnowledgeDiscoveryinDatabase, KDD) 技术中的一个关键步骤
- 提出的背景是解决日益增长的数据量与快速分析数据要求之间的矛盾问题
- 目标：采用各种特定的算法在海量数据中发现有用的数据模式
- 数据挖掘概念
  - 从大量数据中寻找其规律的技术，是统计学、数据库技术和人工智能技术的综合。
  - 从数据中自动地抽取模式、关联、变化、异常和有意义的结构；
  - 大部分的价值在于利用数据挖掘技术改善预测模型。
- 数据挖掘与KDD
  - 知识发现 (KD) 输出的是规则
  - 数据挖掘 (DM) 输出的是模型
  - 共同点
    - 两种方法输入的都是学习集 (learningsets)
    - 目的都是尽可能多的自动化数据挖掘过程
    - 数据挖掘过程并不能完全自动化，只能半自动化
- 数据挖掘的社会需求



- 政府
- POS.
- 人口统计
- 生命周期
- 事实
- 关系
- 模型
- 关联规则
- 序列
- 贸易选择
- 在哪儿做广告
- 销售的地理位置

## 数据爆炸, 知识贫乏

- KDD技术通常包括以下步骤:
  - 理解应用背景
  - 数据准备
  - 数据挖掘
  - 结构解析
  - 使用所发现的知识
- 从KDD的一般步骤来看, 数据挖掘是其中最为关键的处理步骤。所涉及的技术领域知识包括统计学、机器学习、模式识别和数据库技术等
- 技术分类
  - 预言 (Predication) : 用历史预测未来
  - 描述 (Description) : 了解数据中潜在的规律
- 数据挖掘技术
  - 关联分析
  - 序列模式
  - 分类 (预言)
  - 聚集
  - 异常检测
- 与入侵检测相关的算法
  - 主要包括以下三种类型:
    - 分类算法
    - 关联分析算法
    - 序列分析算法
  - 数据挖掘在入侵检测中反复执行
  - 数据挖掘中的关联分析和序列分析算法主要用在模式发现和特征构造的步骤上, 而分类算法主要用在最后的检测模型中
  - 特征的自动构造工作, 体现了数据挖掘在试图解决传统入侵检测中繁重的“知识工程”问题, 即人工编码检测规则问题过程中的某种程度的努力, 目标是建立可扩展和自适应的入侵检测技术

### 异常检测

- 异常检测是数据挖掘中一个重要方面, 用来发现“小的模式”(相对于聚类), 即数据集中显著不同于其它数据的对象
- 异常检测应用
  - 电信和信用卡欺骗
  - 贷款审批
  - 药物研究
  - 气象预报
  - 金融领域
  - 客户分类
  - 网络入侵检测
  - 故障检测与诊断等
- 什么是异常 (outlier) ?
  - Hawkins(1980)给出了异常本质性的定义: 异常是在数据集中与众不同的数据, 使人怀疑这些数据并非随机偏差, 而是产生于完全不同的机制
  - 聚类算法对异常的定义: 异常是聚类嵌于其中的背景噪声。
  - 异常检测算法对异常的定义: 异常是既不属于聚类也不属于背景噪声的点。他们的行为与正常的行为有很大不同。
- 为什么需要预处理
  - 数据
    - 不完整
    - 含观测噪声
    - 不一致
    - 包含其它不希望的成分
  - 数据清理通过填写空缺值, 平滑噪声数据, 识别删除孤立点, 并解决不一致来清理数据。
  - 污染数据形成的原因

- 滥用缩写词
- 数据输入错误
- 数据中的内嵌控制信息
- 不同的惯用语
- 重复记录
- 丢失值
- 拼写变化
- 不同的计量单位
- 过时的编码
- 含有各种噪声
- 数据清理的重要性
  - 污染数据的普遍存在，使得在大型数据库中维护数据的正确性和一致性成为一个及其困难的任务。
  - 垃圾进、垃圾出
- 数据清理处理内容
  - 格式标准化
  - 异常数据清除
  - 错误纠正
  - 重复数据的清除
  - 恢复丢失数据
- 异常检测方法的分类
  - 基于统计 (statistical-based)的方法
  - 基于距离(distance-based)的方法
  - 基于偏差(deviation-based)的方法
  - 基于密度(density-based)的方法
  - 高维数据的异常探测

## 分类算法

- 目标：将特定的数据项归入预先定义好的某个类别。分类算法通常最终生成某种形式的“分类器”，例如决策树或者分类规则等。对于入侵检测而言，理想情况为：
  - 能够收集大量的反映用户或者进程活动的“正常”和“异常”状态的审计数据
  - 选用某个特定的分类算法，经过训练学习生成一个对应的“分类器”
  - 对于输入的先前未见过的新审计记录，该分类器能够准确识别数据项属于“正常”还是“异常”类别
- 常用的算法包括：RIPPER、C45、NearestNeighbor

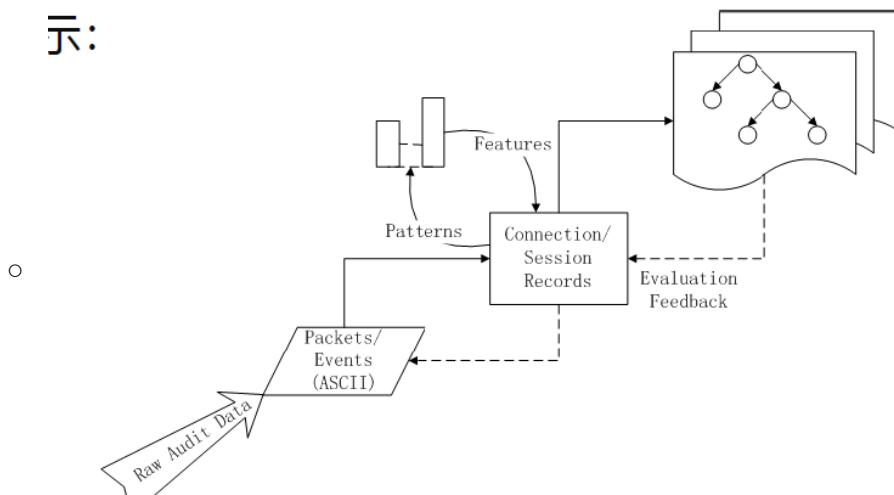
## 关联分析算法

- 用于确定数据记录中各个字段之间的联系
- 入侵检测可以采用这些关联分析算法对审计数据中各个系统特征进行关联分析，例如用户审计数据中命令字段和参数字段之间的关联情况，从而可以用来建立起正常用户行为档案

## 序列分析算法

- 发掘数据集中存在的序列模式，即不同数据记录间的相关性。序列分析算法能够发现按照时间顺序，在数据集合中经常出现的某些审计事件序列模式
- Lee等人最早将数据挖掘引入到入侵检测的领域，并系统提出了用于入侵检测的数据挖掘技术框架。数据挖掘技术应用到入侵检测的技术流程如图所示：

示：

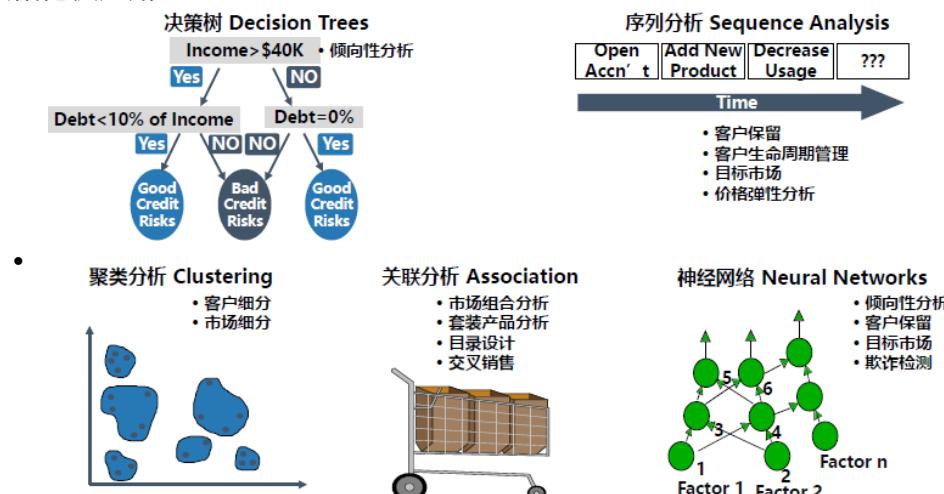


- 如图所示，数据挖掘流程为：
  - 原始审计数据被处理成为ASCII码形式表示的网络数据包信息或者是主机事件数据
  - 这些数据信息进一步被转换成为面向网络连接的记录或者是面向主机会话的记录数据
  - 将数据挖掘的算法应用到这些连接记录上，并计算出有用的数据模式
  - 对这些数据模式进行分析后用来帮助提取连接记录中的其他有用特征
  - 在确定记录数据的特征集合后，应用数据挖掘中的分类算法，对记录数据进行归纳学习，生成最终的检测模型

数据挖掘技术是如何用于构建入侵检测模型——RIPPER分类算法

- RIPPER算法是Cohen于1995年提出的用于数据挖掘的通用分类规则生成算法
- 目标：对输入的数据项进行分类识别
- 每一个RIPPER规则都包含一组条件和一个结论
  - RIPPER规则的条件通常指定了对数据项中某个特征值的测试
  - 规则的结论通常指定了数据项所属特定类别的标识
- RIPPER的学习过程分为
  - 成长阶段：RIPPER规则按照一定的评估标准，不断增加规则的条件数目
  - 修剪阶段：RIPPER规则同样按照一定的标准，删除多余的规则条件
  - 最后生成的RIPPER规则同时兼具简洁性和分类准确的特点
- 采用RIPPER规则作为入侵检测模型具备以下优点：
  - 具备if-then格式的RIPPER规则集合，比较直观易于理解，容易结合到现有的基于规则的检测系统中
  - 有利于进行检测规则的人工验证工作
  - RIPPER规则具备良好的概括归纳能力，对于处理已知攻击手段的变种类型或者新的攻击类型，具备较好的分类性能
  - RIPPER规则具备简洁的条件集合，有利于实时入侵检测工作
- 采用RIPPER算识别异常网络攻击行为和正常网络流量的方法：
  - 进行数据的预处理
  - 再应用RIPPER分类规则生成算法
  - 在测试数据集中，如果当前网络连接的属性特征满足某个RIPPER规则的条件，但是指向的服务端口却不符合RIPPER规则的结论，则指示当前网络连接为异常会话的连接。此时的RIPPER规则集合就变成了一个异常检测器

数据挖掘应用

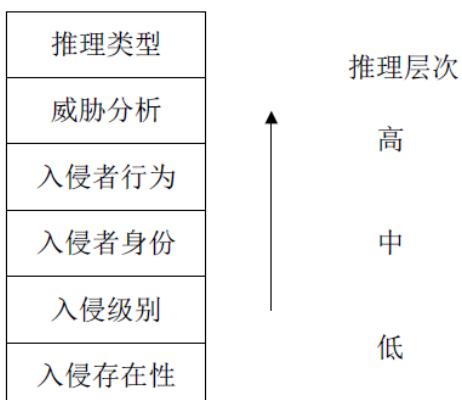


- 数据挖掘成功案例
  - 电话收费和管理办法
  - 竞技运动中的数据挖掘
  - 数据挖掘技术在商业银行中的应用
  - 网上书店关联销售

## 4 数据融合与入侵检测技术

- 数据融合（DataFusion）的概念于20世纪70年代提出，首先在军事领域内得到了应用，到了20世纪80年代末期引起了广泛的重视并被运用到各个研究领域
- 定义（Walatz和JamesLlinas）：

- 数据融合是一种多层次的、多方面的处理过程，这个过程是对多源数据进行检测、结合、相关、估计和组合以达到精确的状态估计和身份估计，以及完整、及时的态势评估和威胁估计
- TBass提出了入侵检测中数据融合的一般层次模型，如下图所示：



- 入侵检测数据融合技术同样面临着若干挑战：
  - 如何开发通用的结构化“元语言”来描述入侵检测和网络管理的对象
  - 对动态网络行为的攻击检测技术
  - 将具有强烈数学背景的多传感器数据融合理论应用到实际的IDS系统所面临的若干复杂问题等

## 5 计算机免疫学与入侵技术

- 计算机免疫技术是直接受到生物免疫机制的启发而提出
- 由于计算机网络受到安全策略、计算机程序以及系统配置等多种因素中所可能包含的错误的影响而总是处于易受入侵的状态，所有入侵检测技术必须要面对这种现实情况
- 与现有的计算机安全系统相比较，生物免疫系统具备如下重要的特征：
  - 多层次保护机制
  - 高度分布式的检测和记忆系统
  - 多样化的个体检测能力
  - 识别未知异体的能力

## 6 进化计算与入侵检测技术

- 对生物进化过程的仿生学研究工作，促进了进化计算（Evolutionary Computation）技术领域的发展
- 进化计算技术本质上属于一种模仿某些自然规律的全局优化算法，其思想可以追溯到达尔文的“进化论”思想，包括自然选择和适者生存的观点
- 进化计算的主要算法：
  - 遗传算法 (Genetic Algorithm, GA)
  - 进化规划 (Evolutionary Programming, EP)
  - 进化策略 (Evolutionary Strategies, ES)
  - 分类器系统 (Classifier Systems, CFS)
  - 遗传规划 (Genetic Programming, GP)
- 进化算法通常维护一组对象的群体，这些对象按照一定的选择规则和遗传算子不断进行进化演变
- 群体中的每个对象个体都具有一个反映其本身于所处环境之间适应性的度量值
- 遗传算子中复制操作主要用于那些具备较高适应性的个体，以发掘个体中蕴藏的适应性特征
- 重组和变异等操作，则用于对个体特征进行适当的扰动，以进行启发性的适应性搜索过程
- 进化算法都可以用于入侵检测，但目前主要是遗传算法和遗传规划，二者的区别：
  - 遗传算法中构成群体的对象个体主要是具有固定长度的字符串或者是比特组
  - 遗传规划中群体的构成个体通常是可执行的程序