文章编号：1005-8451(2007)09-0010-03

采用数据挖掘的入侵检测技术研究

陈 娟，周家纪
（成都理工大学 信息工程学院，成都 610059）

摘 要：基于数据挖掘的入侵检测系统由于其采用数据挖掘技术，成为下一代网络安全防护研究的热点。在数据挖掘技术中，聚类分析是一种重要技术，作者将信息熵理论应用到入侵检测聚类问题中，并通过以部分精度换取聚类高效实现的方式，结合统计的手段实现网络数据包信息的高效聚类。

关键词：入侵检测；数据挖掘；聚类分析；网络技术
中图分类号：TP311 文献标识码：A

Study on intrusion detection based on data mining

CHEN Juan, ZHOU Jia-ji
(College of Information Engineering, Chengdu University of Technology, Chengdu 610059, China)

Abstract: IDS by data mining was becoming a hot topic for network security defense of next generation because of the data mining technology. Cluster analysis was an important means in data mining. It was applied theory of information entropy to the clustering problem for intrusion detection, combined the statistical methods to accomplish a high efficiency purpose.

Key words: intrusion detection; data mining; cluster analysis; network technology

随着网络技术的日益增长，网络问题受到越来越多的关注，传统的静态安全模式已经不能适应新的网络环境，入侵检测系统作为一种主动的安全防护技术，成为传统计算机安全技术的补充。但是由于网络安全本身的复杂性，特别是对于大中型网络系统中巨大的数据处理量问题，目前依然没有一个理想的解决之道，数据处理成为网络安全中急需解决的重要问题，网络安全防护仍处于薄弱状态。

收稿日期：2007-01-19
作者简介：陈 娟，在读硕士研究生；周家纪，教授。

采用数据挖掘相关理论的网络安全技术针对网

地图上点击目标就可以完成登录，即便是对经纬度一无所知的人也可以立即上手。此外，与常用的建筑物登录相比，只增加了经纬度登录，并且用户无需自行输入，而只用鼠标拖拽点击地图就可以完成操作。用户无需任何“再学习”即可使用该工具，因此，可以考虑进行实际运用。

4 结束语

为了测试本系统在实际中的应用情况，笔者请16名志愿者试用了本系统，并在试用后对每个志愿者进行了问卷调查。结果显示，对“本系统使用方便”一项的认同率为100 %，其中理由为“不需要输入”的占56 %，“操作步骤少”的占88 %。

实际测试结果表明，利用方位信息技术能够减少手机末端的输入和操作，提高检索速度。另外，尽管所有志愿者都没有接触经纬度信息的经验，但结果所有人都成功进行了登录，可见使用方位信息登录工具，对使用者并不构成知识障碍，任何人都能够进行方位信息登陆。

参考文献：

[1] 坪井貴彦.携帯電話を用いた建設現場向け作業管理システム[D].平成18年度東洋大学大学院工学研究科情報システム専攻修士論文.
[2] アイティーブースト.はじめてのJSP&サーブレットプログラミング改訂[M].(第3版) 東京：株式会社秀和システム，2005.
[3] 山田祥寛.10日でおぼえるJSP/サーブレット入門教室[M].(第2版) 東京：株式会社翔泳社，2004.
[4] 志村伸弘.MySQL徹底攻略ガイド[M].東京：株式会社技術評論社，2002.

络中海量数据，可以有效提高数据处理效率，解决数据瓶颈问题，成为网络安全领域研究中的一个新热点。

1　传统入侵检测系统

面对越来越严重的危害计算机安全的种种威胁，入侵检测系统（IDS）可以弥补防火墙等传统安全防范手段的不足，为网络安全提供实时的入侵检测及采取相应的防范手段，它在不影响网络性能的前提下，能对网络环境进行监测，从而提供对内部进攻、外部攻击和误操作的实时保护。一个合格的入侵检测系统可以大大地简化管理员的工作，并保证网络安全的运行。

目前，入侵检测系统在异常检测方法上仍存在不足：（1）无法收集到所有的正常数据集，使得当检测到正常数据集中没有、但又是正常行为的数据时，容易误认为是攻击数据，这样误报率将更高；（2）很难保证收集到的正常数据集里不含有攻击记录，如果正常数据里含有攻击数据，并误认为是正常数据进行学习，那么在以后的检测中将会把攻击记录当作正常行为而产生漏报。

入侵检测系统相对于其他静态安全模式的安全防范手段而言，虽然具有较大的优势，但是随着网络规模的不断扩大，传输速率的成倍提高，以及新型入侵手段的不断出现，入侵检测技术面临极大挑战。数据挖掘技术可以帮助人们在海量数据信息中提取出有效的入侵信息，抽象出有利于进行判断和比较的特征模型，降低检测系统的误报率，提高处理实时性。

2　基于数据挖掘的入侵检测

数据挖掘是一个比较新型的研究领域，即从数据中发现肉眼难以发现的固定模式或异常现象。数据挖掘遵循基本的归纳过程，将数据进行整理分析，并从大量数据中提取有意义的信息和知识。与传统基于预定义检测模式的入侵检测技术不同，基于数据挖掘的入侵检测系统可以自动地从训练数据中提取出用于入侵检测的知识和模式，并具有检测效率高、自动化程度高、自适应能力强以及虚报率低等优势。

2.1　数据挖掘与入侵检测融合的可行性

目前，网络中监测到的数据种类繁多，数据量非常大，入侵检测系统具有稳定的数据来源，非常适合用于数据挖掘；通过入侵检测系统在网络中监听到的数据按其所具有的不同性质可以进行分类，同时，不同数据之间的确存有某种相关性，一种连接往往伴随另一种连接的发生。因此，运用数据挖掘技术对审计数据进行挖掘可以得到有价值的信息；从各种渠道所获得的审计数据经过加工处理后，适合运用数据挖掘中的关联分析方法。

2.2　基于数据挖掘的分布式入侵检测模型

基于数据挖掘的分布式入侵检测模型DADIDS（Data mining and Agent-based Distributed Intrusion Detection System），是将数据挖掘应用到分布式入侵检测系统中而建立起来的一种入侵检测模式，其结构模式如图1所示。

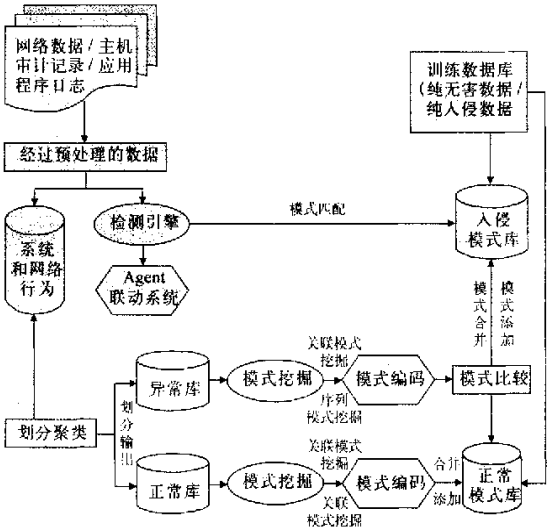

图1　入侵检测结构模式

在基于挖掘的分布式入侵检测系统中，首先将收集到的纯净数据集和有害数据集进行模式挖掘，形成基础的正常模式库和入侵模式库；再将实际环境中搜集到的网络和主机活动进行预处理，形成系统和网络行为集，利用聚类的划分方法对其进行处理，初步区分出异常行为和正常行为，形成异常行为库和正常行为库。对形成的异常行为库和正常行为库进行关联规则挖掘和序列模式挖掘，对正常行为库挖掘出来的关联模式和序列模式经过模式编码后添加到正常模式库；对异常行为库挖掘出的关联模式和序列模式与正常模式进行比较，确定出入侵

模式加入到入侵库中。基于基础库根据实时收集到的数据，利用聚类等数据挖掘手法，不断地对正常数据库和异常数据库进行补充，可以使系统能够构建完整的检测库，并具有实时分析能力。

2.3 入侵检测中的聚类挖掘技术

聚类是将物理或抽象的集合分组成为由类似的对象组成的多个类的过程，由聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异。聚类挖掘是一种无指导的学习方法，它不需要纯洁的训练数据，而是以相似性为基础将记录分组，并依据异常检测的假设来确立是正常模型或异常模型，从而可以很好地解决传统检测方法中存在的问题。

利用聚类算法进行数据的分类，首先需要解决初始聚类核的问题。所谓初始聚类核就是在聚类的最初先确定将要对数据进行大约K种的分类，随后对数据集内每一条数据逐条与其他数据通过公式（1）进行信息熵计算，记录下每条数据的最小熵随后进行降序排序。

$$E(T)=E(X_1)+E(X_2)+\cdots+E(X_n) \quad (1)$$

公式（1）中，$E(X)=-\sum_{X\in S(X)}p(x)\log(p(x))$，$S(X)$表示由属性X进行分类后得到的分类集合，$p(x)$表示每个分类集合占整个数据的比例。

选取队列中的前K个元素作为分类的初始核。在完成初始核的选取工作后将剩余数据逐条聚类到每个分类中，通过将欲分配记录分别到K个不同类中计算信息熵期望：$E(D')=\sum_{i=1}^{K}\frac{|C_i|}{|D'|}E(C_i)$，其中$|D'|$表示已经分类了的数据总数，$|C_i|$表示聚类$i$的记录条数，$E(C_i)$表示聚类$C_i$的信息熵，选取信息熵最小期望值，将其记录加入到分类集合中。

2.4 一种新的聚类处理方式

上述算法在数据较少时可以较快地实现数据的聚类分类，但当数据集合大量时，在各个“数据雪球”滚动到一定程度后算法效率将大大降低，很难适应实时性要求，原因在于每处理一条记录就要对各个分类进行一次信息熵计算，随后进行一次信息熵期望计算，需要进行K次数据库扫描，并在每次扫描中对每个分类的每个属性进行一遍信息熵计算，这在数据量很大时计算量是惊人的；导致系统处理速度非常缓慢。

本文提出了一种新的聚类处理方式。首先根据系统处理性能指定一个适当的阈值，当被处理完了的数据量没有达到指定的阈值时，系统可以提高较高精度的数据聚类处理，否则将采用另一种基于统计的聚类处理方式来对后续数据进行处理。具体算法如下：

对每个分类进行一次扫描，统计各个属性A_i的每种可能取值A_{ij}的出现次数d，并求出每个A_{ij}在这个分类中所有取值数中所占的比例$P(A_{ij})$，$P(A_{ij})=d/e$，e为分类i中属性的所有取值数量，即A_{ij}出现的次数。对于每一个类构造一个统计表，用于记录各个属性的$P(A_{ij})$，然后再对将要进行聚类的每一条记录；对照各个分类结构的统计表分别计算此记录中各个属性的$P(A_{ij})$求和并进行比较，并将此记录加入到获得最大值的分类中。对于无法确定其归属的记录，则可以对各个属性事先赋予不同的优先级，当最大$P(A_{ij})$相同时则根据优先级决定记录归属。

这种算法可以提供对记录的不精确聚类，它牺牲了一定精确度，极大地提高了聚类的速度。利用聚类的挖掘方法研究基于网络的入侵检测行为，可以为系统提供较高精度的正常数据库和异常数据库，降低误报率和漏报率。

3 结束语

由于传统的基于防火墙、身份认证以及加密技术的网络安全防御体系本身存在缺陷和不足，使得入侵检测技术成为当前网络安全方面研究的热点和重要方向，改变了以往被动防御的特点，能够主动实时地跟踪各种危害系统安全的入侵行为，并作出及时响应，成为继防火墙后又一道主动安全防线。在当前的入侵检测技术中，基于数据挖掘的入侵检测技术有较好的发展前景。它将数据挖掘技术引入到入侵检测系统中，在智能性、准确性和扩展性方面得到很大提高。

本文在入侵检测技术有关数据挖掘的聚类算法研究中，在基于信息熵理论的聚类算法基础上，提出了在数据量很大的情况下利用统计分析的手段对聚类挖掘进行处理，通过降低部分精度来获取聚类算法的高效执行。

参考文献：

[1] 桂云苗，朱金福. 一种用信息熵确定聚类权重的方法[J]. 统计与决策，2005（8）.