基于数据挖掘的网络入侵检测技术:研究综述*)

罗 敏 张焕国 王丽娜

(武汉大学计算机学院 软件工程国家重点实验室 武汉430072)

(中国科学院软件研究所计算机科学重点实验室 北京100080)

A Survey of Data Mining-Based Intrusion Detection

LUO Min ZHANG Huan-Guo WANG Li-Na

(Department of Computer Science and technology, State Key Laboratory of Software Engineering, Wuhan University, Wuhan, 430072)

(Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, 100080)

Abstract Intrusion detection techniques are attracted attention. There are a lot of projects which have their own emphasizes in the aspect now. The basic ideas and techniques of data mining-based intrusion detection and the architecture of a real time data mining-based IDS are discussed Some problems and the future directions are proposed.

Keywords Network security, Intrusion detection, Data mining, Architecture

1 引言

随着网络技术和网络规模的不断发展,网络入侵的风险性和机会也越来越多,网络安全已经成为人们无法回避的问题。因此为了保护现在越来越多的敏感信息,入侵检测技术也成为了一种非常重要的技术,得到了越来越多的重视。

入侵检测技术主要分为两类,即:异常检测(abnormal detection)和误用检测(misuse detection)。异常检测是指利用定量的方式来描述可接受的行为特征,以区分和正常行为相违背的、非正常的行为特征来检测入侵。误用检测是指利用已知系统和应用软件的弱点攻击模式来检测入侵[20]。

对于异常检测,人们主要依赖于他们的直觉和经验选择统计特性以构造入侵检测系统。而对于误用检测,则一般首先是由网络安全专家们对攻击模式和系统弱点来进行分析和分类,然后再手工地建立相应的检测规则和模式来构造入侵检测系统。这样就造成了现有的许多入侵检测系统只能够对某一些特定的或已知的入侵行为取得比较好的结果,而且他们的更新是非常昂贵和缓慢的。为了解决这些问题,出现了一些有效的基于数据挖掘的入侵检测技术,也得到了比原有的入侵检测系统更好的结果。

本文介绍几种基于数据挖掘的入侵检测方法,介绍一种基于数据挖掘的实时入侵检测系统的体系结构,讨论该领域当前存在的几个问题和今后的研究方向。

2 基于数据挖掘的几种入侵检测方法

计算机网络的出现产生了大量的审计数据,而且审计数据大都是以文件形式存放,如果仅仅只依靠手工的方法去发现记录中的异常现象是不够的,往往费时费力。数据挖掘技术的快速发展为解决这个问题提供了一种好的解决模式。在众多的数据挖掘方法中有几类方法对入侵检测非常有用。比如说分类的方法,关联分析的方法以及序列分析的方法等等。下面就介绍几种相关的方法的基本思想。

2.1 基于聚类的方法

所谓聚类(clustering)就是将物理或抽象对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(cluster)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,与其它簇中的对象相异。

而基于聚类的无监督异常入侵检测方法建立在两个假设上。第一个假设是正常行为的数目远远大于入侵行为的数目。第二个假设是入侵的行为和正常的行为差异非常大。该方法的基本思想就是由于入侵行为是和正常行为不同的并且数目相对很少,因此它们在能够检测到的数据中呈现出比较特殊的特性。

L. Portnoy 以及 E. Eskin 等人提出了一种基于聚类的使用无类标识数据的无监督异常入侵检测方法[1,2]。其方法是,使用一个简单的基于距离的度量方法来将数据实例形成簇。因为这个聚类过程是在无类标识的数据上进行的,因此只需要提供没有类标示的特征向量。一旦数据形成了簇,就可以把在小簇中的所有实例定义为异常。这个方法的原理可以用上面所说的两个假设来解释。由第一个假设,因为正常行为的数目远大于入侵行为的数目,就可以知道正常的数据实例相对于入侵来说应该形成更大的簇。由第二个假设,因为入侵和正常行为的差异非常大,他们就不可能被分到同一个簇,所以就可以利用该方法来处理异常入侵检测。

该方法最大的优点就是可以不再需要用人工的或其他的方法来对训练集进行分类,而这种分类一般来说是非常困难并且昂贵的。但是该方法的漏警率比较高,造成这种情况的主要原因是该方法的性能主要依赖于训练集的好坏,也就是训练集中包含的入侵行为和正常行为的种类的多少。

2.2 基于判定树的方法

所谓判定树(decision tree)是一个类似于流程图的树结构,其中每个内部节点标示在一个属性上的测试,每个分枝代

*)基金项目:国家自然科学基金课题(66973034),国家自然科学基金重点课题(90104005)资助。罗 敏 博士生,主要研究领域为网络信息安全。张焕国 教授,博士生导师,主要研究领域为密码学与信息安全理论与技术。王丽娜 博士,副教授,主要研究领域为密码学与信息安全理论与技术。

表一个测试输出，而每个树叶节点代表类或类分布。树的最顶层节点是根节点。判定树归纳的基本算法是贪心算法。

RIPPER(Repeated Incremental Pruning to Produce Error Reduction)是由 William Cohen 开发的一种规则学习算法[3]。Wenke Lee 等人改进了 RIPPER 来使用它学习规则以进行异常检测[4,5]。每一个 RIPPER 规则由多个条件(例如：多个属性值)和一个结果(例如一个类标识)的连接组成。如果一个数据项的属性值满足一个规则的条件，我们就称该数据项被这个规则覆盖。对于一个规则的目标类，属于这个类的数据项被称之为正例，那些属于其它类的数据项被称之为反例。

其基本方法如下：一个 RIPPER 规则的学习分为两个阶段，生长阶段和剪枝阶段。整个训练集 D 也因此分为一个生长集和剪枝集。一个规则依靠重复地增加满足最大化 FOIL 信息获取准则[6]的条件来从一个空连接开始生长，这种重复的增加一直到该规则在生长数据集中没有覆盖任何反例的数据项才会停止。在生长阶段后，一个规则会立即进行修剪(例如简化)，这种修剪的做法是重复地删除那些能够在剪枝数据集中导致一个更准确的规则的条件，简单地说就是使规则的产生更加准确和简洁。随之就产生了相应的 RIPPER 规则。

该方法一般是用来进行分类，由于这种方法产生的规则具有 if-then 标准格式，因此这种规则的匹配需要的计算量比较小，所以比较适合实时的入侵检测系统使用。但是和聚类的方法一样，该方法对训练集的要求很高，理想情况是要求提供所有的可能的正常行为。

2.3 基于神经网络的方法

所谓神经网络简单地说就是一组连接的输入/输出单元，其中每个连接都与一个权相联。在学习阶段，通过调整神经网络的权，使得能够预测输入样本的正确类标号来学习。

A. Chosh 和 A. Schwartzbard 于1999年提出了一种基于人工神经网络的入侵检测方法[7]。其目标是能够从不完整的数据中归纳以及能够将在线的数据分类为正常的和入侵的。其基本思想是：一个人工神经网络由简单处理单元，或节点，以及它们之间的连接组成。两个单元之间的连接有权值，这个权值用来说明一个单元影响另一个单元的程度。网络单元的一个子集作为输入节点，另一个子集作为输出节点。在为每个输入节点指定一个值或者一个赋活(activation)函数，并且允许赋活函数能够通过网络繁殖之后，一个神经网络就可以将一组值(指定给输入节点的)映射到另外一组值(由输出节点得到的)。A. Chosh 和 A. Schwartzbard 实现了一个多层前馈感知器网络来进行入侵检测。在这种类型的网络上后向传播算法是一种最普遍的算法。为了使用这种网络，有五个主要的问题：1)怎样对网络的输入数据进行编码。2)怎样确定网络拓扑结构。3)怎样训练网络。4)怎样利用监督训练算法来完成异常入侵检测。5)怎样处理数据网络产生的数据。限于篇幅，这五个问题的解决方法请参阅文[7]。

利用神经网络的方法来进行入侵检测，因为神经网络的优点是其对噪声数据的高承受能力，以及它对未经训练的数据分类模式的能力，所以其优点在于可以解决目前的一些入侵检测系统的通病：不能够从以前观测得到的行为归纳并且识别出将来有可能发生的类似的行为。而神经网络的缺点是往往需要很长的训练时间，而且需要大量的参数，这些参数通常主要靠经验确定，因此其可解释性很差。另外神经网络还存在着过学习的问题，其理论研究还没有取得突破性的进展。

2.4 基于关联规则的方法

关联规则挖掘的目的就是为了发现大量数据中项集(itemset)之间有意义的关联或相关联系。假设 A 是一组属性，项(item)I 是一组在 A 上的值，I 的任意一个子集称之为项集(itemset)，一个项集中的项的数目称之为它的长度，D 是含有 n 个属性(列)的一个数据库。定义 $support(X)$ 为 D 中包含项集 X 的事务(记录)的百分比。那么一个关联规则就是这样一个蕴含式：$X \rightarrow Y$. $[s,c]$，其中 X 和 Y 是项集，并且 $X \cap Y = \phi$。$s = support(X \cup Y)$ 是规则的支持度，$c = \frac{support(X \cup Y)}{support(X)}$ 是规则的置信度。给定两个阈值，最小支持度阈值(minimum_support)和最小置信度阈值(minimum_confidence)。如果 $support(X) \geqslant minimum\_support$，我们称一个项集项 X 为一个频繁(frequent)的项集。如果一个规则同时满足最小支持度阈值和最小置信度阈值，我们称该规则为强关联规则。

Agrawal 等人于 1994年提出的关联规则的算法 Apriori[8]被 Wenke Lee 等利用在入侵检测中[9]。该算法的基本思想如下：挖掘分为两步，第一步是找出所有的频繁项集。可以证明一个频繁的项集的任意非空子集也一定是一个频繁的项集。因此算法首先从 D 中找到长度为1的频繁的项集，然后从那些长度为 k 的频繁的项集迭代的计算长度为 k+1的频繁的项集。当没有新的频繁的项集生成时迭代终止。根据定义，这些项集出现的频繁性至少和预定义的最小支持度阈值一样。第二步是由频繁项集产生强关联规则。根据定义，这些规则必须要满足最小支持度阈值和最小置信度阈值。整个挖掘过程的总体性能由第一步决定。

该方法的优点是不需要专业知识，算法的实现比较简单。但是同样对训练集的要求很高。

2.5 基于序列模式的方法

序列模式挖掘(sequence pattern mining)是指挖掘相对时间或其他模式出现频率高的模式。序列模式挖掘有几个重要的参数，如时间序列的持续时间，事件重叠窗口和被发现的模式中时间之间的时间间隔。还可以在要挖掘的序列模式上指定约束，方法是提供“模式模板”，其形式可以是系列片段(serial episode)，并行片段(parallel episode)，或正则表达式。

关联规则挖掘中采用的 Apriori 特性可以用于序列模式的挖掘，因为若长度为 k 的序列模式是非频繁的，其超集(长度为 k+1)不可能是频繁的。因此，虽然说考虑的参数设置和约束都有所不同，但是序列模式挖掘的大部分方法都采用了类 Apriori 算法的变种。比如在文[9]中，在 Apriori 算法的基础上，增加了一个事务的发生间隔的约束条件。

2.6 其它方法

由于篇幅所限，还有一些入侵检测方法，例如：基于学习的方法[10]，基于规则的方法[17]，基于 Bayes[18]的方法等等[11~16]这里就不再介绍了。

3 一种基于数据挖掘的实时入侵检测系统的体系结构

Wenke Lee、Salvatore J. Stolfo 等人最近研发了一种基于数据挖掘的实时入侵检测系统(IDS，Intrusion Detection System)[19]。其体系结构如图1所示。

整个体系结构包含了传感器，检测器，响应中心，一个数据仓库以及一个模型生成器组件。这个体系结构不仅支持数据的收集，共享和分析，还能够支持数据的归档以及检测模型的生成和分发。

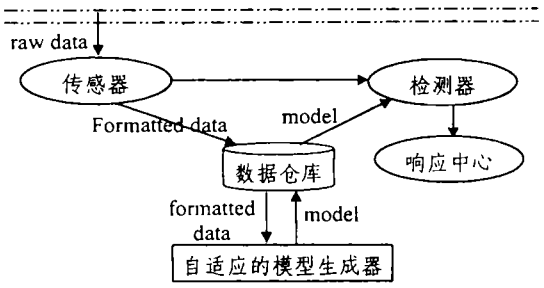

图1 基于数据挖掘的实时入侵检测系统的体系结构

该系统的设计与传感器的数据格式和模型的表示无关。其中每一个数据可以包含任意数量的属性，而每一个属性也可以是连续的或是离散的，可以是数字的或是符号的。在这个框架中，模型可以基于数据挖掘的，例如第2节所提到的基于聚类，判定树，神经网络，关联规则，序列模式等方法，或者是基于规则集的，也可以是基于统计的。为了处理这种多样型，在这个框架中使用了 XML 编码（这种编码能够很方便地转换为CIDF 或 IDMEF（Intrusion Detection Message Exchange Format）格式）以便各个组件间能够很方便地交换数据和模型。该体系结构的最关键的优点在与它的高性能和可升级性。也就是说所有的组件都可以共存在同一个本地网络中，在这种情况下，检测任务被分配到组件中；或者是各个组件也可以在不同的网络中，在这种情况下，它们可以与网络上的其它的入侵检测系统一起合作。

3.1 传感器

传感器的主要任务就是获取原始数据，并且为了在模型评估中使用这些数据对原始数据要进行一些预处理。传感器将入侵检测系统的其余部分和被监视的目标系统的低级属性隔离开来。在该系统中可以利用任何类型的传感器，包括已有的一些网络包的收集工具等等。

3.2 检测器

检测器的主要任务是处理从传感器送来的数据，并且使用一种检测模型来评估这些数据以决定这是否是一次攻击。检测器的另一个任务就是将结果送到数据仓库中留作以后的分析和报告。在同一个系统中可以有多个或多层的检测器来监视系统，并且检测器也可以使用任何一种合适的模型来检测攻击，比如使用第2节所介绍的一些方法产生的检测模型。

3.3 数据仓库

数据仓库主要是用来集中存储数据和模型。这样的一个好处就是不同的组件可以异步地处理存储在一个数据库中的同一块数据，比方说离线的训练和人工的作标记。同一种类型的组件，例如多个传感器也可以同时地处理数据。而相关的数据库的特性也为有效地处理数据提供了方便。

数据仓库也促进了来源于不同的传感器的数据的整合。由于可以利用来源于不同的入侵检测系统的数据和长期以来收集的数据，使得对复杂的和大规模的分布式攻击的检测成为了可能。

3.4 模型生成器

模型生成器的主要任务是加快产生和分发新的（或更新的）入侵检测模型。因为在这个体系结构中，一开始检测结果为异常的攻击可能含有由模型生成器处理的典型数据，所以利用来源于数据仓库的归档的（或历史的）正常的和入侵的数据集，就可以自动地生成能够检测新的入侵的模型并且将它分发到检测器（或任何其它可以使用这些模型的入侵检测系统）中。模型生成器可以看成是一个黑匣子，输入是训练集，输出是检测规则。

3.5 响应中心

响应中心的主要任务是在接到入侵的报警报告后采取相应的措施来处理入侵，比如说关闭端口，断开连接等等。措施的采取可以人工地进行，也可以自动地进行。同时响应中心还需要对报警进行记录，以便日后查询。

3.6 该系统的工作机制

整个系统由上面介绍的五个组件组成，其工作方式如下：首先传感器收集数据并对其进行预处理，使这些数据成为符合系统要求的格式化数据。然后数据将会在检测器中进行检测，看是否有入侵发生，一旦检测器检测到入侵，就可以将入侵报告送到响应中心，由响应中心来决定怎样处理入侵。同时传感器还会将这些数据送进数据仓库，数据仓库组件再将这些数据存放进数据库中。这样模型生成器所需要的训练集就可以利用数据库查询从数据库中得到。模型生成器利用这些训练集可以建立相应的模型，然后再将这些新的或更新的模型分发到各个检测器中。

3.7 该系统的效率问题

效率问题是入侵检测系统必须要考虑的一个重要问题。由系统的工作机制，我们可以看到由于在进行实时检测时，数据只需要在检测器中进行模型的匹配，而匹配过程是非常高效的过程，因此完全能够满足实时检测的要求。另外 IDS 在满足实时检测的同时也应该尽可能小地占用被监控的系统的资源。在基于主机的 IDS 中，后者尤其重要，因为在这种情况下，IDS 往往是直接占用了被监控的主机系统的资源。因此该系统被设计成为一个分布式的系统，模型生成器是一个分开的相对独立的一个系统，这样就可以尽可能小地避免占用被监控的主机系统的资源。而为了减小被监控的主机系统的负担，检测器还可以很容易地从主机系统中移开。在这种情况下，将只有传感器在消耗被监控的主机系统的资源。

结束语 由于已有现成的数据挖掘算法可以利用，基于数据挖掘的入侵检测技术得到了飞速的发展。这种技术的优点是可以处理大量数据的情况，但是它也存在着一些问题。比如，寻找更加高效的数据挖掘算法；入侵检测的正确率；在异常检测中如何能更好地控制误警率；如何能够将该技术应用于实时入侵检测系统中；如何将入侵检测系统和网络管理系统整合在一起；入侵检测技术的形式化设计以及入侵检测过程的可视化问题等等。这些也都是今后的研究方向，都需要大量的研究和实验以开发出更加高效的数据挖掘算法和合适的体系结构。另外，数据挖掘是建立在大样本的基础上的，而对于入侵检测来说，得到足够的样本是非常困难和昂贵的，我们将在这个方面进行一些相应的研究工作。

参 考 文 献

1 Portnoy L, Eskin E, Stolfo S. Intrusion detection with unlabeled data using clustering. In: Proc. of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001), Philadelphia, PA: Nov. 2001

2 Eskin E. Anomaly detection over noisy data using learned probability distributions. In: Proc. of the Intl. Conf. on Machine Learning. 2000

（下转第117页）

进行条件检查的规则。另外，Gaia 模型中的角色权力和职责也融入到这一部分中，这样把传统中和角色紧密相关的一些规则剥离开，提升到了组织规则中来。当我们要全局修改一些协同规则时，可以方便地在规则池中进行添加、删除或修改，不需要对角色或结构进行全局改动，提高了系统的灵活性。

进一步来说，规则虽然约束的是角色间的协同行为，但我们规则池中的规则不涉及具体的角色或行为。这样同一条规则可以被不同的协同行为所指引，对协同进行约束，实现规则的复用。每一个协同行为可以指向多个规则，如果一个角色 Protocol Activity 的协同规则需要进行动态的变动，则只需向规则池中添加没有的规则，并改变与该 Protocol Activity 相关的 Rule Ref. 即可。

规则的描述可使用类似 Tuple Space 和 Linda 描述的方法[3]。

结束语 本文回顾了目前国际上协同和组织研究领域的一些工作进展，并针对我们的协同联盟系统作了一些改进工作，提出了协同系统中组织模型的概念，包括组织角色模型，组织结构，组织规则三部分主要内容，并给出了具体描述方法。

协同联盟组织模型的研究还有很多内容，怎样解决组织的角色和实体(Agents，构件等)的匹配问题，如何使角色更好地适应组织规则，如何重用组织规则，同时也需要加强对协同技术的研究。

参考文献

1 陶先平，吕建. 一种新的移动 Agent 结构化迁移机制的设计和实现. 软件学报，2000，11(7)：918～923

2 吕建，张鸣，廖宇，陶先平. 基于移动 Agent 技术的构件软件框架研究. 软件学报，2000，11(8)：1018～1023

3 Zambonelli F. Coordination Models and Technologies for Internet Agents. In EASSS: Coordination of Internet Agents, August 2000

4 Wooldridge M, Jennings N R, Kinny D. The Gaia Methodology for Agent-Oriented Analysis and Design. Autonomous Agents and Multi-Agent Systems, 2000, 3(3): 285～312

5 Kolp M, Castro J, Mylopoulos J. A Social Organization Perspective on Sofware Architectures. In: Proc. of the First Intl. Workshop From Software Requirements to Architectures (STRAW'01), Toronto, May 2001

6 Yu E. Modelling Strategic Relationships for Process Reengineering: [Ph. D. thesis]. Department of Computer Science, University of Toronto, Canada, 1995

7 Chung L K, Nixon B A, Yu E, Mylopoulos J. Non-Functional Requirements in Software Engineering, Kluwer Publishing, 2000

8 Zambonelli F, Jennings N R, Wooldridge M. Organisational Rules as an Abstraction for the Analysis and Design of Multi-Agent Systems. IJSEKE, 2001, 11(3): 303～328

(上接第107页)

3 Cohen W W. Fast effective rule induction. In: Machine Learning: the 12th Intl. Conf. Lake Taho. CA. 1995. Morgan Kaufmann, 1995

4 Lee W, Stolfo S J. Data mining approaches for intrusion detection. In: Proc. of the 7th USENIX Security Symposium. San Antonio. TX. Jan. 1998

5 Lee W, Stolfo S J, Chan P K. Learning patterns from unix processes execution traces for intrusion detection. In AAAI Workshop on AI Approaches to Fraud Detection and Risk Management. AAAI Press, 1997. 50～56

6 Quinlan J R. Learning logical definitions from relations. Machine Learning, 1990, 5: 239～266

7 Chosh A, Schwartzbard A. A study in using neural networks for anomaly and misuse detection. In: Proc. of the 8th USENIX Security Symposium, 1999

8 Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proc. of the 20th VLDB Conf. Santiago. Chile. 1994

9 Lee W, Stolfo S J. A data mining framework for building intrusion detection models. In: Proc. of the 1999 IEEE Symposium on Security and Privacy. May 1999

10 Chan P K, Stolfo S J. Toward parallel and distributed learning by meta-learning. In AAAI Workshop in Knowledge Discovery In Dtabases. 1993. 227～240

11 Mannila H, Toivonen H, Verkamo A I. Discovering frequent episodes in swquences. In: Proc. of the 1st Intl. Conf. on Knowledge Discovery in Databases and Data Mining. Montreal, Canada. Aug. 1995

12 Ghosh A K, Schwartzbard A, Schatz M. Learning program behavior profiles for intrusion detection. In: Proc. of the 1st USENIX Workshop on Intrusion Detection and Network Monitoring. USENIX Association, April. 1999

13 Ghosh A K, Schwartzbard A, Schatz M. Using program behavior profiles for intrusion detection. In: Proc. of the SANS Intrusion Detection Workshop, Feb. 1999

14 Ghosh A K, Wanken J, Charron F. Detecting anomalous and unknown intrusions against programs. In: Proc. of the 1998 Annual Computer Security Aplications Conf. (ACSAC'98), Dec. 1998

15 Anderson J P. Computer security threat monitoring and surveillance: [Technical Report]. James P. Anderson Co., Fort Washington, PA, April 1980

16 Cannady J. Artificial neural networks for misuse detection. In: Proc. of the 1998 National Information Systems Security Conf. (NISSC'98). Arlington, VA. Oct. 1998. 443～456

17 Porras P A, Kemmerer R A. Penetration state transition analysis - a rule-based intrusion detection approach. In Eighth Annual Computer Security Applications Conf. IEEE Computer Society Press. Nov. 1992. 220～229

18 Schultz M G, Eskin E, Zadok E. Data mining methods for detection of new malicious executables. In: Proc. of IEEE Symposium on Security and Privacy (IEEE S&P-2001). Oakland, CA: May 2001

19 Lee W, Stolfo S J, Chan P K. Real time data mining-based intrusion detection

20 蒋建春，马恒太，任党恩，卿斯汉. 网络安全入侵检测：研究综述. 软件学报，2000，11(11)：1460～1466

21 胡侃，夏绍玮. 基于大型数据仓库的数据挖掘：研究综述. 软件学报，1998，9(1)：53～63

22 Han Jiawei, Kambe M. 数据挖掘概念与技术. 机械工业出版社，2001