第29卷 第16期 计 算 机 工 程 2003年9月
Vol.29 № 16 Computer Engineering September 2003
·基金项目论文· 文章编号：1000—3428(2003)16—0048—02 文献标识码：A 中图分类号：TP393.08

聚类算法在网络入侵检测中的应用

向 继，高 能，荆继武
(中国科学院研究生院信息安全国家重点实验室，北京 100039)

摘 要：分析了目前的入侵检测技术，提出了使用聚类算法进行网络入侵检测的方法，并通过试验说明了该方法的应用效果。
关键词：聚类算法；网络入侵检测；数据挖掘；K-means算法

Application of Cluster Algorithm in Network Intrusion Detection

XIANG Ji，GAO Neng，JING Jiwu
(The State Key Laboratory of Information Security,Graduate School of the Chinese Academy of Sciences,Beijing 100039)

【Abstract】This paper analyses the current intrusion detection techniques, brings forward a technique that applies cluster algorithm to network intrusion detection, and shows the effect through an experiment.
【Key words】Cluster algorithm; Network intrusion detection; Data mining; K-means algorithm

1 基于异常的入侵检测

基于异常的入侵检测技术可以分为需要指导的异常检测和无需指导的异常检测，需要指导的异常检测通过观察得到的正常数据建立正常数据模型，然后检测那些偏离正常模型的异常数据，一个比较典型的使用这种技术的系统是美国乔治梅森大学的MADAM/ID系统。这种方法能够检测新的攻击类型，因为这些新的攻击数据也会偏离正常的数据模型。需要指导的异常检测的一个缺陷是需要一组完全正常的数据来训练获得模型，如果训练数据中包含攻击数据的话，这些攻击就很难检测到，因为该方法把这些攻击数据认为是正常数据，另一方面，要获取这些训练数据也是很困难的。

目前入侵检测技术研究的重点转移到了无需指导的异常检测上，这种技术用一组没有标记的数据作为输入，发现其中存在的攻击数据，即试图从一组不知道什么是正常，什么是异常的数据集中发现那些异常的数据。无需指导的异常检测与需要指导的异常检测相比，它不需要完全正常的训练数据，只需要未加工的网络原始数据。

无需指导的异常检测技术有一个基本的假设，就是正常数据和异常数据有定性的不同，这样才可以将它们区分开来，例如通过一般的分析，可以知道拒绝服务攻击的数据在属性取值和模式上与正常的数据有很大的不同，所以可以利用无需指导的异常检测技术来有效地检测出拒绝服务攻击。

下面介绍的利用聚类算法的异常检测就是一种无需指导的异常检测技术，这种方法可以在未标记的数据上进行，它将相似的数据划分到同一个聚类中，而将不相似的数据划分到不同的聚类，并为这些聚类加以标记表明它们是正常还是异常，然后将网络数据划分到各个聚类中，根据聚类的标记来判断网络数据是否异常。在后面的试验中可以看到，经过聚类后，正常网络数据和异常网络数据被有效地区分。

2 聚类算法简介

聚类算法是一个将数据集划分成若干个聚类的过程，使得同一聚类内的数据具有较高的相似性，而不同聚类中的数据不具有相似性。相似或者不相似根据描述数据的属性值来度量，通常使用基于距离的方法。通过聚类，可以发现数据的密集和稀疏的区域，从而发现数据整体的分布模式，以及数据属性间有意义的关联。

聚类算法涉及很多个领域，包括数据挖掘、统计、机器学习、空间数据库技术，目前研究重点是基于距离的聚类算法。聚类算法也是一种无指导的学习，它不像分类算法那样需要事先标记好的训练数据。

聚类算法的输入是一个包含多个数据的数据集，每个数据通常用一个属性向量$(x_1,x_2,\ldots,x_p)$来表示，其中x_i是一个连续的或离散的变量，代表数据的一个属性的取值。聚类算法的输出是若干个聚类，每个聚类中至少包含一个数据，而且同一个聚类中的数据具有相似性，不同聚类的数据不具有相似性。

为了使用聚类算法，需要计算数据之间的差异，数据间的差异通常用距离来表示，距离计算方法包括欧几里德距离、Manhattan距离、Minkowski距离。其中最常用的是欧几里德距离，它的计算方法如下：

$$d(i,j)=\sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\cdots+|x_{ip}-x_{jp}|^2}$$

其中i,j分别代表数据集中的两个数据，它们都有p个属性。

可以给不同的属性赋以不同的权值，这样计算出来的距离称为加权的欧几里德距离，定义如下：

$$d(i,j)=\sqrt{w_1|x_{i1}-x_{j1}|^2+w_2|x_{i2}-x_{j2}|^2+\cdots+w_p|x_{ip}-x_{jp}|^2}$$

聚类算法有很多种，可以划分为几种类别：划分方法，层次方法，基于密度的方法，基于网格的方法和基于模型的方法，同时聚类也可以用于异常值(outlier)检测。

最常用的聚类算法是K-means算法，它是一种划分方法。给定一个包含n个数据的数据集，和产生的聚类的个数k，K-means算法将n个数据划分成k个子集，每个子集代表一个聚类，同一个聚类中的数据之间距离较近，而不同聚类的数据间距离较远。每个聚类由其中心值来表示，通过计算

基金项目：国家高技术研究发展计划(“863”计划)：信息安全新技术研究——坚固网关技术(2001AA144050)
作者简介：向 继(1976－)，男，硕士生，研究方向：信息安全；高 能，博士；荆继武，教授
收稿日期：2002-09-03

聚类中所有数据的平均值可以得到它的中心值。K-means算法描述如下：

算法：K-means

输入：聚类的个数k和一个包含n个数据的数据集

输出：k个聚类

方法：

任意选取k个数据作为初始的聚类中心

Do

将每个数据划分到一个聚类，依据数据与聚类中心值的距离最近为标准。

更新聚类的中心值，即重新计算每个聚类中所有数据的平均值

While划分没有发生变化

K-means算法有良好扩展性，被广泛应用于各领域。

3 在网络入侵检测中使用聚类算法

网络入侵检测目的是分析网络中传输的数据流量，从中发现异常的流量。目前绝大部分的网络入侵检测系统采用的技术都是基于特征的检测技术，下面提出一种方法，它利用聚类算法来实现网络入侵检测系统，这种方法和基于特征的检测方法相比，在不需要训练数据的情况下就能够检测出新的类型的攻击，其中特别适合于检测拒绝服务、端口扫描等类型的攻击。

运用聚类算法进行网络入侵检测的过程通常可分3个阶段，分别是收集分析数据、对分析数据进行标准化和运用聚类算法对数据分类。

3.1 收集分析数据

在计算机网络系统中，局域网普遍采用的是基于广播机制的IEEE 802.3协议，即以太网协议，该协议保证传输的数据包能被同一冲突域内所有的主机接收，网络入侵检测正是利用这一特性收集网络上传输的数据流量。

通常网络入侵检测系统利用一些抓包工具来收集网络数据包，其中最常用的工具是Unix操作系统下的Tcpdump，它能够监听和接收网络中所有正在传输的数据包，并把它们记录到文件中。

原始的网络数据包本身还不适合于进行聚类分析，需要将原始的网络数据包恢复成TCP/IP层的连接记录，其中每个连接记录代表一次TCP/IP层的连接事件，包含一个网络连接的多个属性，包括网络协议、连接起始时间、连接结束时间、服务(端口号)、源IP地址、目的IP地址、连接终止状态、TCP标志等。现在我们获得了进行聚类分析的数据集——一组连接记录其中每个数据就是一个网络连接记录。

3.2 数据标准化

在运用聚类算法之前，还需要对数据的属性值进行标准化，因为属性值之间的差别可能很大，而且它们可以用不同的单位来度量，例如时间可以用s来度量，也可以用ms来度量，但使用不同的度量方法，对数据间距离的影响也不同。为了消除由于度量不同对距离产生的影响，需要对属性值进行标准化。标准化的方法如下：

首先计算各个属性值的平均值m和平均绝对偏移S。

$$m_j = \frac{1}{n}(x_{1j} + x_{2j} + \cdots + x_{nj})$$

$$S_j = \frac{1}{n}\left(|x_{1j} - m_j| + |x_{2j} - m_j| + \cdots + |x_{nj} - m_j|\right)$$

其中m_j和S_j分别是第j个属性的平均值和平均绝对偏移，

分别是各个数据第j个属性的取值。

然后对每个数据的每个属性进行如下的标准化：

$$z_{ij} = \frac{x_{ij} - m_j}{S_j}$$

其中Z_{ij}代表标准化后的第i个数据的第j个属性。这种转变实际上将待处理的数据从它原来所处的空间转换到一个标准化的空间。

3.3 运用聚类算法

通过收集连接记录和对其进行标准化后，获得了进行聚类分析的数据集，接下来可以利用聚类算法来对这些连接记录进行分类，区分出哪些是正常的连接记录，哪些是异常的连接记录。可以采用的聚类算法有很多，例如可以采用层次的聚类方法和基于模型的聚类方法，但是由于K-means算法具有较小的计算复杂度和良好的扩展性，以及网络入侵检测对于实时性的要求，我们选用K-means算法作为进行聚类分析的算法。

运用聚类算法的结果是产生了若干个聚类，每个聚类中包含部分的连接记录，根据第1节中的假设，正常的连接记录与异常的连接记录有定性的不同，所以它们之间不具有相似性，应该处于不同的聚类之中。这样可以把包含异常连接记录的聚类标记为异常聚类，而将包含正常连接记录的聚类标记为正常聚类。

由于没有使用带标记的训练数据，因此无法知道哪些连接记录是正常的，哪些是异常的，也无从对聚类作出相应的标记，解决这一问题的方法有很多，例如可以作出另一种假设，正常连接记录的数量远大于异常连接记录的数量，在实际的网络环境中，这种假设是可以成立的，这样就可以采用这样的方法标记各个聚类：如果聚类中的连接记录的个数大于某一阈值，就将它标记为正常的聚类，否则标记为异常的聚类。

利用聚类分析产生的结果来检测入侵的方法很简单，对于一个连接记录d，首先对它进行标准化，然后从获得的聚类集合中找到一个聚类C，使它的中心值与d的距离最近，根据C的标记对d进行分类。

4 性能测试

为了验证聚类算法在网络入侵检测中应用的效果，我们利用KDD CUP 99数据集进行了试验。

4.1 试验过程

试验选用的数据集是KDD Cup 99 数据集，它是为进行网络入侵检测试验而由美国国防部专门收集的数据集，它包含多种类型的攻击数据和正常数据，有近500万条连接记录，每个连接记录都被标记为正常或者一种特定类型的攻击。提取的连接记录的属性包括一些描述TCP连接的基本属性，如持续时间、协议类型、传输的字节数、指示正常和错误状态的TCP标记。还有一些基于2s时间窗的统计属性，如到达同一个主机的连接数，有SYN和REJ错误的连接数，属于同一类型服务的连接数。

为了便于试验，我们对数据集进行如下的筛选：集中于检测拒绝服务攻击，将不属于拒绝服务攻击和正常的数据去除，这样数据的类型只包括back，land，neptune，pod，smurf，teardrop和normal 7种，前6种是拒绝服务攻击类型。为了进一步减少数据量，从每个类型的数据中取出至多两万条记录，这样最后得到待分析的连接记录的总数为63470

(下转第185页)

据库中下载相关模型。然后用数据驱动模型生成与A一致的传动机构零件图。B根据传动机构尺寸及定位方式设计箱体，完成后进行装配，当需要螺栓、轴承等标准件时，可通过附加菜单调用服务器端的标准件库管理模块，该模块把相关数据及模型传给B。PRO/E生成所需规格标准件图形用于装配。如果B在装配过程中发现传动机构某零件与其它零件干涉，B将该零件新的约束条件传送给A，A重新设计该零件，并将结果提交给B。B继续装配，直至设计结果满意为止。LDT80提升机的装配模型如图6所示。

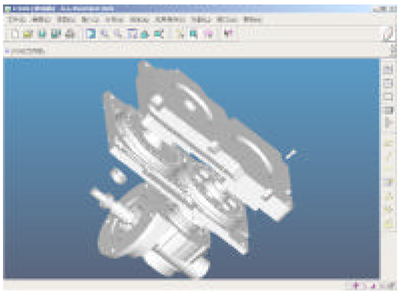

图6 LDT80提升机装配模型图

5 结论

本文分析了设计资源库对产品设计支持的重要性，并研究了若干建库的关键技术，实现了一个基于Web服务的协同环境下设计资源库系统，并用示例验证了该方法的可行性。

通过Web服务的方法可以实现系统集成，有望解决传统设计资源库在异地协同环境下遇到的跨越平台与系统的困难。今后的研究将深入探讨异构CAD系统之间参数化模型的统一，设计资源库与PDM、ERP等的集成。

参考文献

1 Franca G,Marina M,Domenico B.A Modeling Tool for the Management of Product Data in a Co-design Environment.2002,34(14):1063-1073

2 张 威陈定方魏道政.一个机械产品协同设计系统的研究与实现.计算机辅助设计与图形学学报,2000,12(11):864-866

3 张宏展,胡正国.对象Web环境下遗留系统集成技术研究.计算机工程,2002,28(1):20-22

4 孙宏伟,张树生,王 静.组件化松耦合企业应用集成关键技术研究.计算机应用,2002,22(4):4-8

5 余安萍,俞俊平,孙华志.C#程序设计教程.北京电子工业出版社,2002-01

☆☆

(上接第49页)

个。由于原来的记录中属性很多，只选取其中的15个属性，其中包含3个离散的属性和12个连续的属性。这里利用了带标记的连接记录，目的只是为了后面说明聚类的结果。

对筛选后的连接记录利用3.2节中的方法进行标准化，但是只是对取值连续的属性进行了标准化，而对取值离散的属性没有进行标准化。然后采用K-means算法对连接记录进行聚类分析。首先假定输出10个聚类，从数据集中任选10个数据作为这10个聚类初始的中心值，然后将每个数据划分到其中的一个聚类，方法是判断该数据到哪一个聚类的中心值的距离最近，则将该数据划分到这个聚类。划分完成后，重新计算每个聚类的中心值，方法是计算聚类中所有数据的平均值，把这个平均值作为新的聚类的中心值。接下来重新将每个数据划分到其中的一个聚类，如此反复地进行若干次划分和计算中心值的操作，直到数据的划分没有变化。

在运用K-means算法时，我们用欧几里德距离方法计算连接记录之间的距离，但是对于那些取值离散的属性，采用特殊的处理，如x_{im}为一个离散的属性，如果数据i和数据j在该属性上取值不同，用一个常数来代替欧几里德距离公式中的$|x_{im}-x_{jm}|^2$.

4.2 试验结果

进行聚类后，整个数据集被划分成4个聚类，各个聚类中含有的正常数据和攻击数据的数量如表1。

从表1中可以看出，聚类1和聚类4中正常数据占绝大多数，聚类2和聚类3中拒绝服务攻击数据占绝大多数，所以可以标记聚类1和聚类4为正常，而聚类2和聚类3为异常。

表1 聚类划分结果

聚类	正常	拒绝服务攻击
1	2	0
2	14	323
3	144	40213
4	19841	2933
合计	20001	43469

从入侵检测的角度来分析试验结果，分别计算出这种方法的检测率为93.3%，误警率为0.79%，它们分别代表检测为攻击的攻击数据占所有攻击数据的比重，以及检测为攻击的正常数据占所有正常数据的比重。

通过上面的数据分析可以看到，利用聚类算法可以有效地将正常数据和攻击数据区分开来，且具有很好的准确性，特别对于拒绝服务攻击这样的攻击具有很高的检测率和很低的误警率。

参考文献

1 Kdd99 Cup dataset.http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html,1999

2 Portnoy L,Eskin E,Stolfo S J.Intrusion Detection with Unlabeled Data Using Clustering. Philadelphia,PA:In Proceedings of ACM CSS Workshop on Data Mining Applied to Security(DMSA,2001),2001

3 Eskin E,Arnold A,Prerau M.A Geometric Framework for Unsupervied Anomaly Detection: Detecting Intrusions in Unlabeled Data.CUCS Technical Report,2002

4 Application of Data Mining to Intrusion Detection.http://www.isse.gmu.edu/~csis/infs765/handouts/handout12.pdf,2000

5 Barbara D.ADAM:Detecting Intrusions by Data Mining.Proceedings of IEEE Workshop on Information Assurance and Security,2001

6 Han Jiawei,Kamber M.数据挖掘—概念与技术.高等教育出版社,2002