聚类在网络入侵的异常检测中的应用

Application to Cluster Algorithm in Anomaly Detection of Network Intrusion

严晓光 （华中科技大学 机械科学与工程学院 430074）
褚学征 （华中科技大学 软件学院 430074）

摘要：介绍了网络入侵检测技术；分析了数据挖掘中的聚类算法在网络入侵异常检测中的应用，给出了系统的总体模型设计，并分析了各个模块的功能；通过实验结果中验证了该方法的有效性。

关键词：异常检测 数据挖掘 聚类

1 入侵检测技术

由于互联网的复杂性不断增长、网上的服务种类不断增多，使得网络的安全问题越来越突出。从大型的商业网络到小型家庭办公网络，信息安全都受到威胁。要保证网络的安全，直接措施是防止对网络的攻击行为，通常采用防火墙技术。

防火墙一方面可能被攻破或被绕过，另一方面过度使用防火墙会妨碍网络正常运行，因此不能完全依靠防火墙来防止网络入侵。网络系统还需要能及时发现恶意行为，并在这种行为对系统或数据造成破坏之前采取措施，如发出警告、切断连接、封掉 IP，甚至进行反击等，这就是入侵检测技术[1]。入侵检测方法一般可以分为误用检测技术和异常检测技术两大类。

（1）误用检测技术需要事先将入侵信息的特征信息输入入侵信息特征库，检测时依据具体特征库进行判断是否有入侵，准确度很高。但对新的特征库以外的入侵信息难以识别。适用于防范已知特征的入侵信息。

（2）异常检测技术的通用性较强，可能检测出以前未出现过、未知特征的的攻击方法。现在已经成为主要的研究对象。

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，在监测的过程中那些偏离正常数据模型的数据将被视为异常。这种检测的要求是在训练的数据中必须全部是正常的数据；对于现实中网络流量很大、需要分析的数据很多的情形，这一要求很难满足。一旦训练数据中包含了攻击数据，在训练中它就会被当成正常数据，这些攻击就很难被检测到。

无需指导的异常检测技术而是根据正常数据和异常数据不同特征，将它们区分开来。这种技术不需要完全正常的训练数据，用可以用来从源于网络的原始数据中发现存在的攻击数据。

2 数据挖掘技术

数据挖掘（Data Mining）又被称为数据库中的知识发现，是指从大型数据库或数据仓库中提取隐含的、未知的、异常的或有潜在应用价值的信息或模式，是数据库研究中的一个很有应用价值的新领域。数据挖掘是一门汇集统计学、机器学习、数据库、人工智能等学科内容的新兴的交叉学科。根据挖掘功能数据挖掘可以被分为分为：特征化、区分、关联、分类、聚类、孤立点分析、演变分析、偏差分析等等。

聚类是数据挖掘的基本形式之一。按照数据基本特征的相似性和差异性，将数据划分为若干组（组内还可以再分组），同组数据的特征尽量相似，不同组的尽量相异，这种对数据进行自动分组的方法称为聚类[5]。在机器学习领域中聚类是无指导学习的一个例子，与分类不同，聚类和无指导学习不依赖于预先定义的类和带标号的训练实例。所以聚类分析很适用于入侵检测方面的研究。

本文中主要介绍将数据挖掘中的聚类算法应用于

异常检测技术，并利用这种技术设计相关的软件，实现预期的功能。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。根据聚类的思想在网络入侵检测中收集数据，大部分的正常数据聚集在一起成为一大簇。小部分入侵数据聚类之后单独会聚成一小簇。这样，先对数据进行聚类分析，然后将小簇数据标识为入侵数据而得出新的异常行为模式，最后由系统自动转换为检测规则。

3　系统模型设计

根据上面所说的无需指导的异常检测在入侵的异常检测方面的优点，以及聚类技术具有无指导学习的特点，设计出了基于聚类技术的网络入侵异常检测系统的模型如图1。

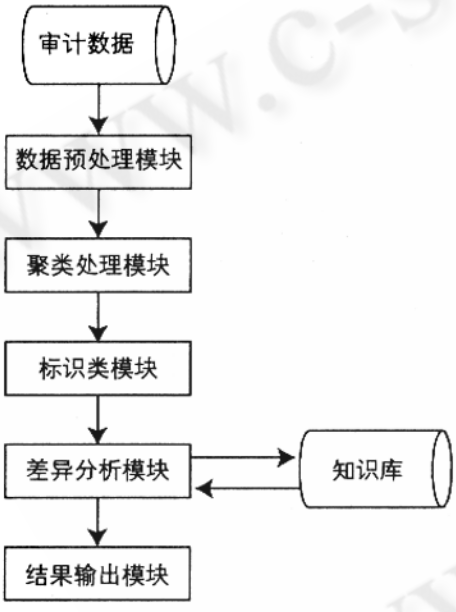

图 1　基于聚类技术的网络入侵异常检测系统模型

3.1 收集审计数据

数据采集是入侵检测的此系统基础，此模块主要是从网络上提取数据进行监测，例如网络层通过的原始 IP 包、链路层的数据帧等。由于在局域网中普遍使用的是 IEEE802.3 协议，主机之间传送数据时采用子网内广播的方式，即任何一台主机向子网内的某台主机发送数据时，该数据均会在其子网内广播，也就是说该数据会被任何一台主机作为数据包接受（只要将网卡设置为混杂模式）。这样的优点是数据的采集不会影响到主机和网络的性能。此系统采用非常成熟的 Unix Tcpdump 抓包工具来收集网络数据包并把它们记录到文件中。

3.2 数据预处理模块

在运用聚类算法之前，需要将原始数据转化成适合挖掘算法使用的格式，过滤和去掉噪声。为此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，例如时间可以用秒来度量，也可以用毫秒度量；使用不同的度量方法，对数据的影响会不同。为了消除由于度量标准不同而产生的影响，需要对属性值进行标准化。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。为后来的分析做好准备。

3.3 聚类处理模块

在收集的数据（在异常检测中主要是连接记录）已经标准化，获得了可以进行聚类分析的数据集后，接下来就可以利用适当的聚类算法来对这些连接记录进行分类，区分哪些是正常的连接记录，哪些是异常的连接记录。在网络级入侵检测中可以采用的聚类算法有很多，例如可以采用分层次的聚类方法，基于模型的聚类方法和基于统计的孤立点分析的方法等等。

3.4 标识类模块

运用聚类算法的结果将产生若干个簇，每个簇中包含部分的连接记录。正常的连接记录与异常的连接记录的特性不同，因此它们之间不具有相似性，应该处于不同的簇之中。这样就可以把包含异常连接记录的簇标记为异常簇，而将包含正常连接记录的簇标记为正常簇。由于正常数据远远多于异常数据，可以根据簇中所包含的数据的多少，并根据相关的孤立点分析技术来识别它们。

3.5 差异分析模块

从已经完成聚类的结果中抽象出其模式，与知识库中已有的训练模式相对比。判断这种模式是否在知识库中已经存在。如果存在或有相似的模式，就可以判定是否有真正的入侵情况，若有入侵时就输出信号给显示输出模块报警。如果在知识库中没有相似的模式，则作为新模式存储到知识库中，形成知识储备，供以后的模式判别使用。

3.6 结果输出模块

从差异分析模块接收信号和数据，显示是否受到

入侵以及数据细节。

此系统主要是针对聚类算法的无指导的学习的特性在网络的入侵检测中的应用而设计的。具有可以处理大数据量，可以实现自主学习，在遇到入侵的情况下及时、准确地报告具体情况等特点。

4 系统实现

4.1 聚类算法实现

聚类算法有很多种，在各个领域得到广泛的应用，可以划分为几种主要类别：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等。

在聚类算法中最经典、最常用到的划分算法是k—mean 算法和 k—中心点算法。k—mean 算法对于"噪声"和孤立点数据很敏感的，少量的该数据就能够对平均值产生极大的影响。用 k－中心点算法不采用簇中对象的平均值作为参照点，可以选用簇中位置最中心的对象，即中心点。这样划分方法仍然是基于最小化所有对象与其参照点之间的相异度之和的原则来执行的。K—中心点算法处理的数据结果簇紧密、效果较好，在处理大数据量时具有可伸缩性和高效性，而且在有"噪声"的情况下仍然可以得到较好的聚类结果。在实际的大流量的网络环境中存在大量的各种数据纪录，不可避免地有"噪声"存在，因此基于 K—中心点算法来实现此系统。

图 2 是 K—中心点算法的流程图，它的基本思路是：首先为每个簇任意选择一个代表对象；剩余的对象根据其与代表的距离分配给最近的一个簇。然后反复地利用非代表对象来代替代表对象，以改进聚类的质量。用一个代价函数来估算聚类结果的质量，该函数度量对象与其参照对象之间的平均相异度[4]。

4.2 性能测试

系统性能测试时采用了 KDD Cup 1999 Data 的数据集。该数据集首先在与 KDD99 同时举办的第三届国际知识发现和数据挖掘工具竞赛上使用，它包含了在军事网络环境中仿真的各种入侵数据。检测网络入侵的软件能使网络免遭各种非授权用户及潜在入侵者的侵扰。

该数据集提供了从某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，其训练数据集包含 5 百万个连接数据，测试数据集包含了 2 百万个连接数据。一个连接就是在规定的协议下、在规定的时间内完成的起始并终止的 TCP 分组序列，这些序列在固定的源 IP 地址与目的 IP 地址之间进行数据传输。每个连接都带一个类标识，或者是正常，或者是某个具体的攻击类型。图 3 给出了该数据集所含的数据记录示例。

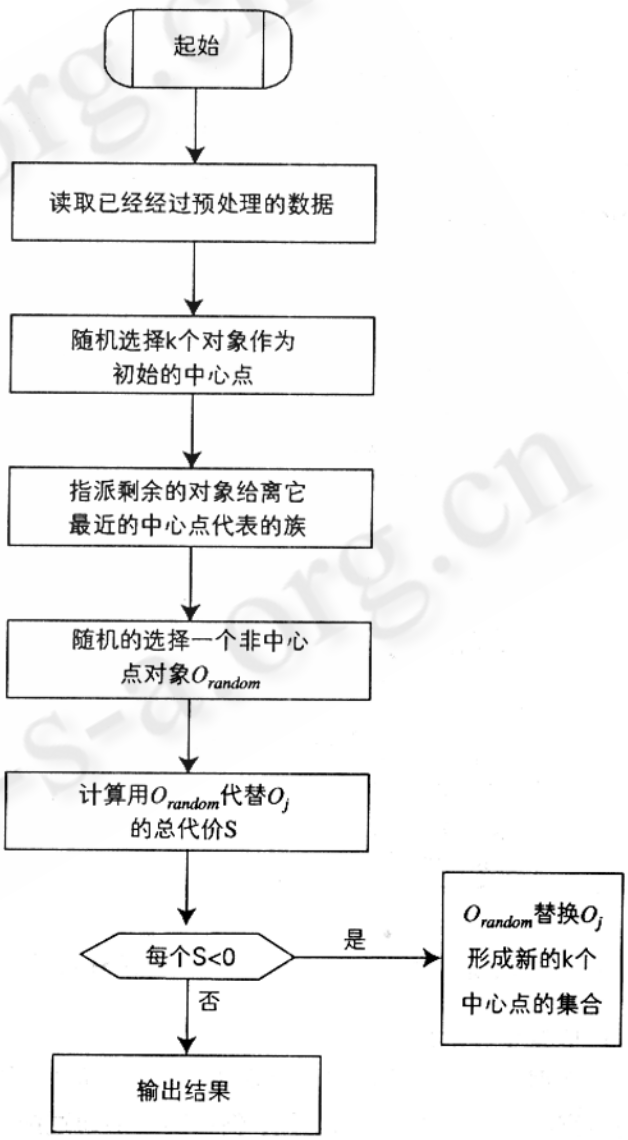

图 2　k—中心点算法的流程图

在上述网络连接中出现的攻击有以下几种类型：

DOS（denial－of service）：拒绝服务攻击，如 ping of death，teardrop，smurf；

282,tcp,ftp,SF,162,597,0,0,….0,0,1,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,warezmaster
0,tcp,ftp_data,SF,12,0,0,0,….0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00, warezmaster
0,tcp,http,SF,282.24572,0,0,….0,0,3,3,0.33,0.33,0.00,0.00,1.00,0.00,0.00,normal
0,tcp,pop_3,SO,0,0,0,0,….0,0,1,2,1.00,1.00,0.00,0.00,1.00,0.00,mscan
0,tcp,telnet,SO,0,0,0,0,….0,0,1.12,1.00,0.25,0.00,0.58,1.00,0.00,mscan
0,tcp,private,REJ,0,0,0,0,….0,0,142,9,0.00,0.00,1.00,1.00,0.06,0.06,0.00,nepture
0,udp,private,SF,105,147,0,0,….0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,snmpgetattack
0,udp,private,SF,105,147,0,0,….0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,normal

图 3　tcp 连接记录示例

R2L:来自远程机器的非法访问,如 guessing password;

U2R:普通用户对本地超级用户特权的非法访问,如各种 buffer overflow 攻击;

Probing:监视和其它探测活动,如 port scanning,ping sweep。

测试中对数据集进行了筛选:主要检测拒绝服务攻击(DOS)和来自远程机器的非法访问(R2L),将不属于实验要求的数据去除。为了进一步减少数据量,从每个类型的数据中取出至多五万条记录,这样最后得到待分析的连接记录的总数为 106370。为了更方便地应用聚类方法来对 TCP 连接进行分类,主要选择连续型属性进行考察,共选择了 14 种属性其中包括 3 个离散的属性和 11 个连续的属性。其中的三个离散型属性主要是为用于对原始数据集进行子集划分以改善算法性能而考虑的。

表 1　聚类后的结果

聚 类	Normal	DOS	R2L
1	22142	1171	1317
2	2579	1533	23972
3	20557	1048	1103
4	3465	26510	973
总计	48743	30262	27365

将选出的数据作为待分析的审计数据,用上面设计的系统进行分析,聚类分析采用 4.1 节介绍的 k—中心点算法。表 1 为经过聚类后的测试结果,从表中可以看出,经过聚类后的实验数据很好地被标示出哪些簇是入侵的数据,哪些是正常数据。可以明显地看出聚类 1 和 3 中正常的数据占的比例很高,说明这两个簇为正常。拒绝服务攻击(DOS)和远程机器的非法访问(R2L)分别在聚类 2 和 4 中占绝大多数,所以聚类 2 是远程机器的非法访问的入侵,聚类 4 是拒绝服务攻击。最后计算出这种方法的检测率为 87.6%,误警率为 0.82%。

5　结束语

本文介绍了数据挖掘中的聚类算法在网络入侵中的异常检测中的应用,并给出了系统的总体模型设计,分析了各个模块的功能,并在实验的结果中说明了该方法的有效性,聚类算法可以有效地将入侵数据和正常数据区分开,并且在检测率和误警率上也取得了较好的效果。

参考文献

1 Herve Debar, Marc Dacier, Andreas Wespi. Towards a taxonomy of intrusion—detection systems [J]. Computer Network, 1999:31:805—822.

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

3 Wenke Lee ,et al . Algorithms for Mining System Audit Data[C]. Proceedings of ,IEEE Symposium on Security and Privacy ,1999.

4 J. Han, M. Kamber, Data Mining :Concepts and Techniques[M]. Morgan Kaufmann Publishers , Inc. ,2004:234 -235S.

5 王能斌,数据库系统原理[M],电子工业出版社,2000。