文章编号:1000 - 582X(2004)03 - 0046 - 03

网络入侵的聚类算法研究与实现*

叶 芳,吴 中 福,刘 勇 国
(重庆大学 计算机学院,重庆 400030)

摘 要:入侵检测中对知入侵的检测主要由异常检测完成,传统的异常检测方法需要构造一个正常行为特征轮廓的参考模型,但建立该特征轮廓和确定异常性报警的门限值都比较困难,而且建立该特征轮廓使系统开销大。据此本文提出一种针对入侵检测的聚类算法和一种数据处理方法。该算法通过动态更新聚类中心和类内最大距离实现,收敛速度快,再结合对数据的预处理使聚类效果更好。实验结果表明,此算法用于以未知入侵检测为代表的特殊模式检测方面是可行和有效的。

关键词:入侵; 网络入侵检测; 聚类
中图法分类号:TP393.08; TP301.6 **文献标识码**:A

随着互联网络的飞速发展,网络信息安全正日益得到人们的关注,入侵检测则成为安全专家积极研究的重要课题。Heady[1]将入侵定义为"任何企图破坏资源完整性、保密性和可用性的行为"。其检测方法主要分为误用检测和异常检测两类,它们各有自己的优势,在不同的安全策略中有不同的应用。但由于入侵类型的日益复杂,新的入侵行为层出不穷,使未知入侵的检测显得尤为重要,而误用检测不能对未知入侵进行模式匹配,因此未知入侵检测主要由异常检测来完成。

目前的异常检测主要是通过构造一个关于系统正常行为的参考模型(有关用户、系统关键程序等的特征轮廓),然后检查系统的运行情况,若与给定的参考模型存在较大的偏差,则认为系统受到了入侵攻击,如 ISA - IDS 系统[2]采用了统计模型,在样本集中对每一个特征进行统计后找出一个中心值,再选择一个偏离门限,只要发生的事件超过这个门限,就被认为是入侵。这种方法设计简单,但检测率不高,适用于具有简单分布的集合。于是有研究人员利用机器学习思想,通过已标识网络数据训练学习算法来检测入侵行为[3-5]。但这种方法必须有学习过程,而学习过程给系统带来很大开销,耗时也较多。据此,笔者提出了一种聚类算法和相应的数据预处理方法用于入侵检测. 通过实验,表明此算法在未知入侵检测方面是可行和有效的,并极大地提高了入侵检测率,同时有效控制了误检率。

1 聚类原理

1.1 入侵检测分析

入侵检测首先是基于两个基本假设:1)用户和程序行为是可见的; 2)正常行为与入侵行为本质上是可区分的。在一般网络环境中,正常行为是主流,而入侵表现为个别现象,正常实例的规模远大于入侵行为数目。由此算法要针对一般网络环境下正常实例规模远大于入侵实例的情形,于是入侵检测的又一个重要假设前提是检测数据模式空间中绝大部分模式属于正常行为,由这个重要特征可考虑将检测数据集分为由正常行为特征的聚类团与各种非正常行为的模式或小聚类团两个类组成,即只需将正常行为模式划分到一个聚类集中,而对各种类型入侵行为模式不作进一步聚类分析,这样建立一个正常行为模式的聚类团,把正常行为模式与入侵行为模式尽量分离. 从这种思想出发,提出了一种针对入侵检测的聚类算法。

1.2 聚类原理

希望经过聚类算法把大部分属于同一类的模式区分出来. 首先,把所有检测数据都看作是正常行为模式(即所有模式属于同一类),计算初始类心,通过给定的参数确定类内最大距离后,调整各模式的类别,判断各模式是否属于正常行为模式类,把不属于该类的模式划分出去,再重新计算类心和类内最大距离,继续判断和调整模式类别,最终找到大部分模式所属的类,实现正常行为模式与入侵行为模式的尽量分离。

这种方法的聚类中心和类内最大距离都是动态变化的,虽然每一次变化都更加接近类的真实情况,但由于算法本身对类的分布的要求有局限性,所以在聚类前,还要对模式进行预处理,使算法适应更多分布情况的类. 该动态聚类原理框架如图 1 所示:

* 收稿日期:2003 - 10 - 08
基金项目:国家自然科学基金资助项目(60271019);重庆市应用基础研究资助项目(7370)
作者简介:叶芳(1977 -),女,重庆人,重庆大学硕士研究生,主要研究领域为入侵检测,神经网络。

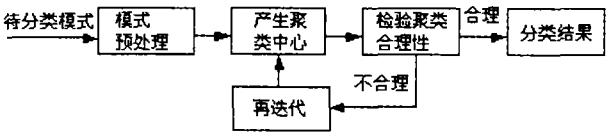

图 1　聚类原理框架图

2　聚类算法实现

2.　条件及约定

设待分类的输入模式特征矢量集 X 为 $\{X_1, X_2, \cdots, X_N\}$，其中大多数模式属于一类，类的数目是 2（正常行为类和入侵行为类）。

2.2　基本思想

首先视所有输入模式为一类（正常行为类），选取 u 值得到初始类内最大距离（各模式到均值的偏离程度），然后通过不断计算类心和类内距离，实现不断地调整模式的类别，直到属于正常行为类的各模式到类心的距离都满足小于类内最大距离的一半为止。

2.3　算法步骤

1）将所有输入模式都看作是正常行为模式，即所有模式属于同一类，确定初值 $k=0$，$X^{(k)}=X$，$N^{(k)}=N$，给出 $u(u>1)$ 值。

2）计算模式集 $X^{(k)}$ 的类心 $C^{(k)}$ 和各模式到类心的平均距离 $S_C^{(k)}$，求 $P^{(k)}uS_C^{(k)}$。

$$C^{(k)}=\frac{1}{N^{(k)}}\sum_{i=1}^{N^{(k)}}X_i^{(k)},\ S_C^{(k)}=\frac{1}{N^{(k)}}\sum_{i=1}^{N^{(k)}}\| X_i^{(k)}-C^{(k)}\|,$$

其中 $X_i^{(k)}\in X^{(k)}$

式中 $N^{(k)}$ 为 $X^{(k)}$ 中所含模式个数，上角标 k 表示迭代次数，$P^{(k)}$ 是阈值.

3）对模式集 $X^{(k)}$ 中的模式逐个求出到类心 $C^{(k)}$ 的距离 $d_i^{(k)}$，并且按 $d_i^{(k)}$ 与 $P^{(k)}$ 的比较情况，判断该模式是否属于 $X^{(k)}$。即：

若 $d_i^{(k)}\leqslant P^{(k)}$，则判 $X_i^{(k)}\in X^{(k)}$；

若 $d_i^{(k)}>P^{(k)}$，则判 $X_i^{(k)}\notin X^{(k)}$ 其中 $i\in N^{(k)}$。

把不属于 $X^{(k)}$ 类的模式 $X_i^{(k)}$ 从该类中划分出去，于是产生新的聚类 $X^{(k+1)}$。

4）如果 $X^{(k+1)}=X^{(k)}$，则结束；否则，$k=k+1$，转至 Step2.

2.4　性能分析

该聚类法是在大多数模式同属于一类情况下的一种的动态聚类法，它的特点是以确定的 u 值和各模式到类心的平均距离的乘积得到类内最大距离为前提，各模式按最小距离原则指定类别，并不断调整，使属于同类的大部分模式被尽量准确地区分出来. 其方法简单，耗时很少，反复实验的聚类结果尚令人满意，如模式分布呈现类内团聚状，算法可达很好的聚类结果。在实际应用中需要探试不同的 u 值，以进一步达到更大范围的聚类效果。

2.5　实例

二维的模式集分布如图 2 中的（a）所示，大部分模式分布在 $\{(x,y)|0.2<x<0.5, 0.2<y<0.5\}$ 中，用上述算法进行聚类分析，选 $u=1.7$，聚类情况如图 2 中（b），圆圈表示大多数模式所属类的范围，每一个圆圈是一次迭代，最大的一个圆是迭代第一次后的情况，最小的圆是聚类完成时的情况，可见类和类的范围不断被修正，各模式类别的指定不断调整，模式集最终被分为了两类。

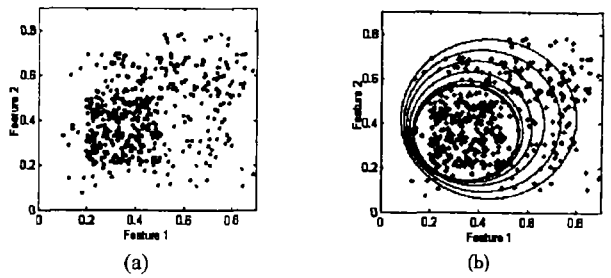

图 2　模式分布与聚类过程

在图 3 中显示了一种特殊情况，即当输入模式集包括多个聚类集时的聚类情况，大部分模式分布与图 2 的情况类似，但在点（0.7，0.6）附近有一个小聚类团如（a）中所示，模式集更难区分，它的聚类分析情况显示迭代次数比图 2 的增多（图 3（b））。因此只要输入模式集中大部分模式同属一类，即使还包含其他小聚类团，也不会影响该算法的聚类效果。

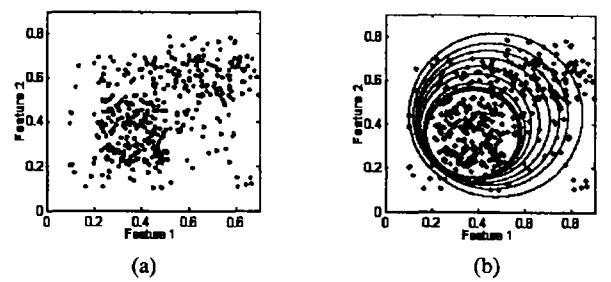

图 3　模式分布与聚类过程

3　数据预处理

3.1　数据集特征分析和预处理

上述聚类算法实现的实质是用一个聚类中心和一个类内最大距离来描述一个类，显然这样的一个类在空间中是呈球状分布的，但在实际应用中类的分布多种多样，如类的分布是长条状时，分类效果就会较差. 所以，在进行聚类算法前，如已知各类模式分布的某些知识后，利用它们来指导聚类，效果就将会提高。针对这一问题很多研究者提出了各种解决方案，如利用类核函数等，而笔者提出的是在保持数据的拓扑结构不变的情况下，把数据在 n 维空间中映射成超球状分布，使类的分布变得更加适合该聚类算法。

现以二维空间为例说明该预处理思想，保持数据拓扑结构不变，将图 4 中的图形 A 映射成图形 B（正方形），再把 B 映射成图形 C（圆形）。从 A 到 B 可通过坐标的伸缩变换得到，即将长方形或多边形映射成正

方形，而从 B 到 C 的映射相当于把正方形映射成它的内切圆，图形可平移成图 5 的情况，设圆心为坐标轴原心，半径长为 r，其映射思想是：先将正方形看作是无数条从 O 点到正方形边沿点的直线组成，如直线 OA 和 OE 等，再伸缩变换直线，如把直线 OA（长为 l）映射成直线 OF，可通过伸缩变换 r/l 得到，这样直线上各点依然保持拓扑结构不变，无数条直线同时伸缩变换可使正方形映射成它的内切圆。

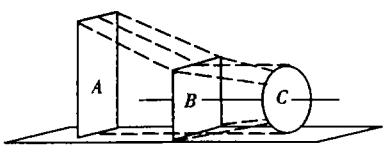

图 4 模式集转化

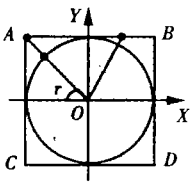

图 5 模式集转化

在按图 5 建立的坐标系中，以 (x,y) 表示原坐标值（原正方形内各点），(x',y') 表示转换后的坐标值（圆中的各点），其正方形映射成它的内切圆的转换公式为：

$$\begin{cases} x' = x \times \dfrac{\max\{|x|,|y|\}}{\sqrt{x^2+y^2}} \\ y' = y \times \dfrac{\max\{|x|,|y|\}}{\sqrt{x^2+y^2}} \end{cases} \tag{1}$$

在 n 维空间中用 $(x_1,x_2,\cdots x_n)$ 和 $(x'_1,x'_2,\cdots x'_n)$ 分别表示原坐标值和转换后的坐标值，则转换公式为：

$$x'_k = x_k \times \frac{\max\left\{|x_k|,\sqrt{\sum_{i=1,i\neq k}^{n} x_i^2}\right\}}{\sqrt{\sum_{i=1}^{n} x_i^2}} \quad k \in \{1,2,\cdots n\} \tag{2}$$

同一类的各个模式之间总存在某些联系，如拓扑结构、特征分布等，从不同角度分析可得到不同的结果，由此根据不同模式识别的需要选择恰当的分析角度，该预处理方法就是将类的分布映射成更适合聚类算法的分布情况，而各模式之间的关联并没有改变，仍然属于同一类。通过这种数据预处理方法来提高聚类算法的效果，使算法不仅仅适合类的分布呈球状的数据，同时结合各类模式分布的具体知识指导数据预处理，扩大聚类算法的适用范围。

4 网络入侵检测实验

入侵检测的数据预处理：实验数据集为 KDD Cup 1999 网络连接数据集[6]，此数据集是 1998 年在麻省理工学院 Lincoln 实验室由 DARPA 举办的为入侵检测模型评估而建立的测试数据集. 每个实例包含 41 个属性，均已标识. 根据该数据样本中的特征数目，可以确定数据空间维数是 41，在 41 维空间中用上述数据预处理方法对其进行预处理。

入侵检测实验结果及分析：选 10 组不同检测数据集，每组包括 5 万条检测数据，分别对每组数据进行入侵检测实验，反复实验试探 u 最恰当的取值，记录其检测结果（图 6）。

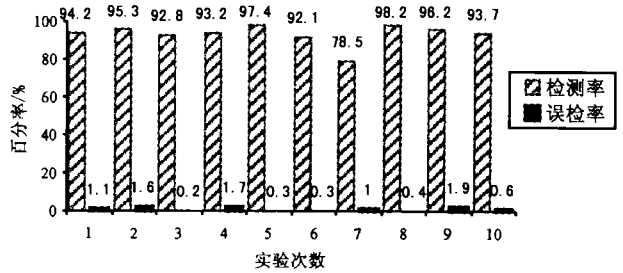

图 6 入侵检测结果

检测率和误检率是 IDS 最重要的性能指标，在图 6 中通过检测率和误检率来表达入侵检测的实验情况。检测率与误检率总是紧密相关，增加检测率常常要以误检率的增加为代价，而误检率偏高使系统对原本不是攻击的事件产生了错误的警报，将导致 IDS 的功效降低. 因此，既能增加检测率又能降低误检率是 IDS 最希望达到的目标。从图中可以看到误报率控制在 2% 以下时，检测率可达 90% 以上，与目前最好的 IDS 的误检率 2% ~3% 时检测率 60% ~70% 相比，有明显提高。但同时也存在一些问题，比如仍然有某些分布很特殊的数据集不能得到很好的聚类效果，还需进一步改进，但这种算法思想作为一种基本的聚类方法是值得推广的。

5 结 语

笔者提出的聚类法是在大多数模式同属于一类情况下的一种新的动态聚类法，是针对于以入侵检测为代表的一类特殊模式分布特征的模式识别，再结合对模式预处理的方法，可使该算法的聚类效果更好。

参考文献：

[1] 孙即祥. 现代模式识别[M]. 北京：国防科技出版社. 2002.

[2] 李家春，李之棠. 入侵检测系统[J]. 计算机应用研究，2002，(11)：5 - 9.

[3] HEADY R, LNGER G, MACCABE A, et al. The Architecture of a Network Level Intrusion Detection System[A]. Technical Report CS90 - 20, Department of Computer Science[C], University of New Mexico, August 1990.

[4] LEE W. A Data Mining Framework for Constructing Features and Models for Intrusion Detection Systems[PhD Thesis]. Columbia University, USA, 1999.

[5] PUKETZA N J, ZHANG K, CHUNG M, et al. A Methodology for Testing Intrusion Detection Systoms[J]. IEEE Transactions on Software Engineering, 1996,22(10):719 - 729.

[6] LEE W, STOLFO S J, MOK K. Data Mining in Work Flow environments: Experience in Intrusion Detection[A]. CHAUDHURI S ed. Proceedings of the 1999 Conference on Knowledge Discovery and Data Mining[C]: ACM Press, 1.

（下转第 52 页）

5 结束语

移动 IP 是基于网络层解决移动问题的方案，当 IPSec 与它结合后，可使它用在 VPN 的移动用户问题上。采用 VPDN 的移动用户通过 NAS(即 LAC)提供与 VPN 内部通讯的隧道连接，但 VPDN 对于通过局域网而非拨号接入的移动用户存在局限性。移动 IP 技术中移动用户可自主的发起与 VPN 中的 GW(即 LNS)的连接。这对于移动用户采用局域网的接入方式与 VPN 通讯是非常方便的。这种模型对于自己建立 VPN 的团体或组织来说比 VPDN 更方便，因为用户不用事先申请 VPDN 业务，不用交纳 VPDN 的费用。而且只要与 VPN 协商好，移动用户可采用任意的隧道，如 IP/IP、GRE 等，因此具有更好的自主性和独立性。

参考文献：

[1] RFC 2685, Virtual Private Networks Identifier[S].
[2] RFC 2003, IP Encapsulation Within IP[S].
[3] 何宝宏. VPN技术综述[J]. 中国数据通信，2002，4:10－14.
[4] RFC 2764, A Framework for IP Based Virtual Private Networks[S].
[5] RFC 2002, IP Mobility Support[S].
[6] RFC 2344, Reverse Tunneling for Mobile IP[S].

Application of Mobile IP in VPN

LUO Ya, YAO Jia-ning, HUANG Fu-tao, ZOU Zong-hui
(College of Computer Science, Chongqing University, Chongqing, 400030 China)

Abstract: VPN is a technology that realizes the security transmission of private information through the public network. As one of key technologies, tunnel technology resolves the mobility problem of mobile nodes. Firstly, L2TP and IPSec have been compared based on these two tunnel principles introduction of VPN. Secondly the advantages of IPSec in security have been introduced. Then, the problems of mobile host and two models of tunnel in VPN have been introduced. Lastly, the advantages of utilizing the mobile IP to resolve the problem of mobile host have been analyzed, and this technology has been compared with VPDN.

Key words: VPN; mobile IP; VPDN; tunnel technology; L2TP; IPSec

(编辑 张 苹)

(上接第 52 页)

Clustering Detection Algorithms for Network Intrusions

YE Fang, WU Zhong-fu, LIU Yong-guo

(Department of Computer Science, Chongqing University, Chongqing 400030, China)

Abstract: Traditional abnormal detection methods need a reference model with a profile of normal action, but building the character profile and specifying threshold of abnormal alarm are difficult. So this paper puts forward intrusion detection in combination with clustering and data processing. This algorithm comes true dynamicly updating the center of cluster and the biggest distance within cluster with fast convergence. The effect is better with the help of pre-processing the data. By means of simulated experiments, this algorithm is proved feasible and efficient for unknown intrusion detection.

Key words: intrusion; network intrusion detection; clustering

(编辑 吕赛英)