基于模糊数据挖掘技术的入侵检测算法与应用

高　翔[2]，王　敏[1]，郭　英[1]
(1. 西北工业大学 计算机学院，陕西 西安 710072;2. 空军工程大学 电讯工程学院，陕西 西安 710077)

摘　要:基于数据挖掘技术的入侵检测技术是近年来研究的热点,目前有不少入侵检测系统中都采用了关联分析的数据挖掘方法,现有的关联分析算法只能够解决数据中分类属性的挖掘,对于数值属性则不能直接使用,然而网络流量数据中包含了许多反映入侵状况的数值属性,已有学者提出了将数值属性先进行分类而后再进行关联分析的挖掘方法,然而这种方法带来的问题是在进行异常和正常划分时存在明确的界限,即"尖锐边界问题",由于网络安全概念自身具有一定的模糊性,因此明确的界限可能会导致误报和漏报的情况产生,从而影响检测效果,文中提出了一种基于模糊关联挖掘技术的入侵检测算法,并采用遗传算法确定划分模糊集合的隶属度函数参数,最后的实验结果说明了该算法的有效性。

关键词:数据挖掘;入侵检测;模糊逻辑;遗传算法;关联分析
中图分类号: TP393.08　　**文献标识码**:A　　**文章编号**:1009-3516(2006)06-0068-04

目前,很多入侵检测系统中都采用了数据挖掘方法来建立用户的行为框架进行网络行为异常检测[1-5]。由于模糊逻辑提供了一种有效地将概念模式进行分类的方法,因此,将数据挖掘和模糊逻辑相结合可以从大量的数据中发现更为广泛的内容,而且相对于单纯数据层次上的挖掘,模糊挖掘可以提取出更高层次的规则,使挖掘结果更容易被用户了解。将模糊逻辑应用到入侵检测中的原因主要有两个:首先,入侵检测中包含了许多数值属性的特点,例如,来自同一个源主机的不同 TCP/UDP 服务数量等,这些都可以看作是潜在的模糊变量;其次,网络安全概念本身具有模糊性。对于具有数值属性的特征,如果将正常值的取值范围设定为一个区间,则任何在此区间以外的都被视为异常,而不管它们和这个区间的差距有多大。这样的结果会让正常和异常之间产生很明显的界限,导致"尖锐边界问题",引入模糊概念可以消除这个界限。

1　模糊关联规则挖掘

如果在一个关联规则的前、后项中包含了用于限定属性的模糊数,则称这样的关联规则为模糊关联规则。给定一个数据库 T 及属性 I,以及与 I 相关的模糊集,就可以找出其中潜在有用的规则,模糊关联规则形式为:X *is* $A\rightarrow Y$ *is* B。

在上面的规则中,$X=\{x_1, x_2,\cdots x_p\}$ 和 $Y=\{y_1, y_2,\cdots y_q\}$ 都是项目集 I 的子集,且 $X\cap Y=\varphi$,$A=\{f_{x_1}, f_{x_2}, \cdots, f_{x_p}\}$ 和 $B=\{f_{y_1}, f_{y_2}, \cdots, f_{y_p}\}$ 包含了分别与 X,Y 相关的模糊集合。

同普通关联规则一样,如果一个规则是有价值的,那么它将具有足够的支持度和置信度。$\langle X,A\rangle$ 支持度 $S_{<X,A>}$ 的计算公式如下:

$$S_{<X,A>}=\frac{\sum_{t_i\in T}\prod_{x_j\in X}\{\alpha_{a_j}(t_i[x_j])\}}{\text{total}(T)},\quad \alpha_{a_j}(t_i[x_j])=\begin{cases}m_{a_j\in A}(t_i[x_j]) & \text{if } m_{aj}\geqslant\omega\\ 0 & \text{otherwise}\end{cases}$$

上面的等式中,$\langle X,A\rangle$ 代表模糊项目集对,X 是属性的 x_j 集合,一个记录满足 $\langle X,A\rangle$ 意味着记录的影响大于

收稿日期:2006-04-05
基金项目:国家自然科学基金资助项目(60573101)
作者简介:高　翔(1974-),男,陕西西安人,博士(后),主要从事计算机软件及网络安全技术研究.

0。记录影响的计算是通过隶属函数来实现的,隶属等级不小于用户给定阈值 ω,过低的隶属等级将不予考虑。用 $t_i[x_j]$ 来代表第 i 个记录中的 xj 值,通过隶属函数计算 $t_i[x_j]$ 值对应模糊集合的隶属等级 。在获得了一条记录中的每个 x_j 对应模糊集合的隶属等级后,使用 $\prod_{x_j \in X}\{m_{a_j \in A}(t_i[x_j])\}$ 来计算记录 t_i 的影响,累加了所有记录的影响后,再除以记录的总数,就可以计算出 $\langle X,A\rangle$ 的支持度。

置信度的计算公式如下:

$$C_{<<X,A>,<Y,B>>} = \frac{\sum_{t_i \in T}\prod_{z_k \in Z}\{\alpha_{c_k}(t_i[z_k])\}}{\sum_{t_i \in T}\prod_{x_j \in X}\{\alpha_{a_j}(t_i[x_j])\}}, \quad \alpha_{c_k}(t_i[z_k]) = \begin{cases} m_{c_k \in c}(t_i[z_k]) & m_{c_k} \geqslant \omega \\ 0 & \text{otherwise} \end{cases}$$

其中 $Z = X \cup Y, C = A \cup B$。置信度反映出支持 $\langle X,A\rangle$ 的项目集中同时也支持 $\langle Z,C\rangle$ 的项目集所占的比例。

模糊关联规则的挖掘算法与布尔关联规则的挖掘算法 Apriori 基本一样,只是在 $k \geqslant 2$ 时,从 L_{k-1}(所有长度为 $(k-1)$ 的频繁项目集的集合)中构建 C_k(长度为 k 的候选频繁项目集的集合)的算法存在区别。

2　隶属函数的构造

必须结合具体的数据来构造合适的隶属函数,实验数据来自 http://iris.cs.uml.edu:8080,我们从该站点下载了 3 组 tcpdump 形式的网络流量数据。其中,base 代表正常状态下的网络流量,net1 是包含了模拟 IP spoofing 攻击的网络流量,在该数据文件中,入侵者试图通过猜测 IP 序列号来获得对远端主机的访问,net3 是包含了模拟端口扫描攻击的网络流量,在该数据文件中,入侵者试图收集有关网络上的主机及其提供的服务的信息。哥伦比亚大学的学者 Lee 和 Stofol 经过实验发现网络流量中的一些数值属性可以用来分析发现入侵。我们采纳了他们的结果,对流量数据进行了预处理,得到以下 3 组数值属性:

S_N:过去 2 s 内,在包头中包含同步标志 SYN 的 TCP 包的数量;

F_N:过去 2 s 内,在包头中包含连接结束标志 FIN 的 TCP 包的数量;

R_N:过去 2 s 内,在包头中包含复位标志 RST 的 TCP 包的数量;这些数值属性计算时采用了宽度为 2 s 的区间重叠。

以上 3 个数值属性可以被看作是模糊变量。我们用 FuzzyCLIPS 提供的 3 个标准隶属函数 S,P_I,Z 分别将这 3 个模糊变量分为 HIGH,MEDIUM,LOW 3 个模糊集合。如图 1 所示。

3 个标准函数 S,P_I,Z 的定义为

$$S(x,a,c) = \begin{cases} 0 & x \leqslant a \\ 2\left(\frac{x-a}{c-a}\right)^2 & a < x \leqslant \frac{a+c}{2} \\ 1-2\left(\frac{c-x}{c-a}\right)^2 & \frac{a+c}{2} < x \leqslant c \\ 1 & c < x \end{cases}$$

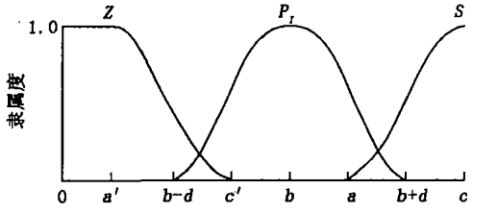

图 1　模糊集合的标准函数表示

$$Z(x,a',c') = 1 - S(x,a',c')$$

$$P_I(x,d,b) = \begin{cases} S(x,b-d,b), x \leqslant b \\ Z(x,b,b+d), b \leqslant x \end{cases}$$

这样问题的关键就是如何确定每一个数值属性对应模糊集合的隶属函数的参数。我们采用遗传算法来确定立书函数的参数,模糊集定义上的每个标准隶属函数都需要确定 2 个参数,因此定义模糊变量的隶属函数总共需要 6 个参数,组成的基因的数据结构如表 1 所示。一条染色体由一组基因构成,每个基因都代表了一个模糊变量。染色体的数据结构如表 2 所示。

表 1　基因的数据结构

Z	a'	c'
P_I	b	d
S	a	c

表 2　染色体的数据结构

S_N		F_N	R_N	P_N
a'	c'			
b	d			
a	c			

遗传算法中适应度函数的定义基于从不同审计数据集中挖掘的规则的相似度。假设 R_1,R_2 分别为规则

集 S_1 和 S_2 中的规则,且形式为:$R_1: X \rightarrow Y, s, c$;$R_2: X' \rightarrow Y', s', c'$。则 R_1 和 R_2 的相似度为

$$\text{Similarity}(R_1, R_2) = \begin{cases} 0 & (X \neq X') \vee (Y \neq Y') \\ \max(0, 1 - \max(\frac{|c - c'|}{c}, \frac{|s - s'|}{s})) & (X = X') \wedge (Y = Y') \end{cases}$$

两个规则集 S_1、S_2 的相似度为:$\text{Similarity}(S_1, X_2) = S_T / |S_1| |S_T / |S_2|$

$|S_1|$、$|S_2|$ 分别代表 S_1 和 S_2 中的规则总数。

使用 3 个不同的审计数据集来测试算法:一个正常的不包含入侵的数据集 base;两个分别包含不同类型入侵的异常数据集 net1,net3。正常数据集划分为两个集合。一部分是 base 数据的 80%,作为参考数据,剩余 20% 作为正常测试数据。遗传算法的适应度函数定义为:$F_1 = S_{rn}/S_{ra1} \cdot S_{rn}/S_{ra2}$

S_{rn} 是挖掘参考数据得到的规则集和挖掘正常测试数据得到的规则集之间的相似度,S_{ra1} 和 S_{ra2} 则分别是挖掘参考数据得到的规则集和挖掘两组异常数据得到的规则集之间的相似度。为避免被零除,以上的适应度函数都在分母里加上一个极小的常数。因为参考数据集的规则集和两组异常数据的规则集之间的相似度必须同时最小化,所以函数 F_1 促使系统对两种类型入侵的灵敏度联合改变,针对所要解决的具体问题,我们设定群体规模为 30,群体初始化时需满足的条件为:$a < c$ *and* $a' < c'$ *and* $c < b < a'$,设定固定交叉的概率 0.6。

3 实验结果

我们采用 F_1 作为遗传算法的适应度函数来确定划分 S_N、R_N、F_N3 组数值属性的模糊集的隶属函数,而后进行模糊关联规则挖掘并分别计算 base20.rul、net1.rul、net3.rul 和参考数据规则集 base80.rul 之间的相似度,设定的阈值为 min_sup = 0.05,min_conf = 0.3,threshold = 0.2,其中 base20.rul,net1.rul,net3.rul 分别代表挖掘正常测试数据和 net1,net3 数据后得到的规则集,结果如图 2 所示。同样的实验数据,采用单纯的关联规则挖掘方法得到的检测结果如图 3 所示。

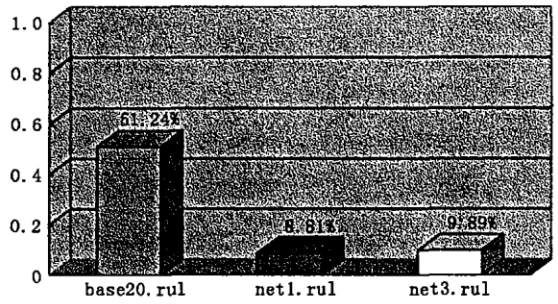

图 2 采用模糊关联规则方法挖掘的规则集相似度计算结果

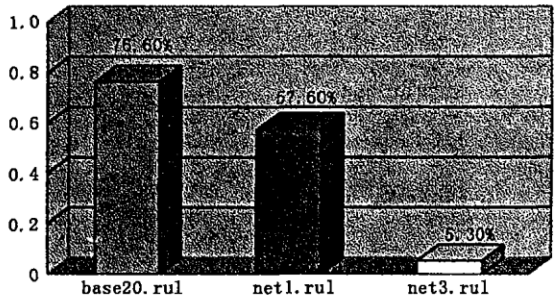

图 3 采用关联规则方法挖掘的规则集相似度计算结果

根据实验结果我们不难发现,结合模糊逻辑的关联规则挖掘方法较好地解决了数值属性的分类,能够发现隐藏在数值属性中的入侵知识,从而取得了较好的检测结果。

参考文献:

[1] 高 翔,王 敏,胡正国. 基于数据挖掘技术的入侵检测系统研究[J]. 西北工业大学学报,2003,21(4):395 - 396.

[2] Han J,Kamber M, 数据挖掘概念与技术[M]. 北京:高等教育出版社,2001.

[3] 高 翔,王 敏,胡正国. 神经网络在异常检测中的研究与应用[J]. 仪器仪表学报,2002,23(3):469 - 470.

[4] 高 翔,苏广文,胡正国. 入侵检测系统中的网络监测[J]. 微电子学与计算机,2002,19(2):37 - 39.

[5] Eugene H Spafford. Diego Zamboni. Intrusion Detection Using Autonomous Agent[J]. Computer New York,2000,34(4):547 - 570.

(编辑:门向生)

The Application of Fuzzy Data Mining Algorithm in Intrusion Detection

GAO Xiang[1], WANG Min[2], GUO Ying[1]
(1. Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China; 2. The Telecommunication Engineering Institute, Air Force Engineering University, Xi'an 710077, Shaanxi, China)

Abstract: Intrusion detection system is a newly emerging and promising security measure. Data mining methods have been used to build automatic intrusion detection systems based on anomaly detection. The goal is to characterize the normal system activities with a profile by applying mining algorithms to audit data so that abnormal intrusive activities can be detected by comparing the current activities with the profile. This paper provides a new Intrusion Detection method based on data mining technology and combines fuzzy logic with apriori mining method. By grouping the quantitative attributes in network traffic according to fuzzy set, and by using genetic algorithm to construct the membership functions that state the fuzzy set, the existing "sharp boundary" problem can be avoided if the classic set theory is adopted. The experiment result shows that this combining fuzzy logic data mining method is an effective anomaly detection way.

Key words: data mining; intrusion detection; fuzzy logic; genetic algorithm; association analysis

(上接第67页)

Abstract: By improvement of Zheng's authenticated key agreement protocol, an identity sign - cryptic technique - based authenticated key agreement protocol is proposed. This protocol has the advantage of sign - cryptic technique and achieves the two functions of authentication and encryption in a single logical step. Therefore, it is of high efficiency. Moreover, due to using ID - based public key system, the expense of building and managing public key infrastructure is decreased and the users need not store or transfer public keys and certificate. And again, in our proposed protocol the bilinear pairing on elliptic curve is employed to reach the equivalent security levels with short length key and small computation cost. As a result, the remarkable properties of our authenticated key agreement protocol are low computation cost, narrow bandwidth requirement, and high security level.

Key words: cryptography; authenticated key agreement protocol; sign - cryptograph; ID - based public key system; bilinear pairing