模糊关联规则算法在入侵检测中的应用

费淑芳 郑 宁 余日泰 徐 明
(杭州电子科技大学计算机学院 浙江 杭州 310018)

摘 要 提出了一种结合 Apriori和 Kuok's算法的改进的模糊关联规则算法。在定义隶属函数、决策树结构和规则集相似度的基础上,采用改进的挖掘算法挖掘数值属性的关联规则。实验结果表明,算法在规则生成和时间效率方面都显示了良好的性能。

关键词 入侵检测 数据挖掘 模糊逻辑 关联规则

APPLYING FUZZY ASSOCIATION RULES ALGORITHM IN INTRUSION DETECTION

Fei Shufang Zheng Ning Yu Ritai Xu Ming
(College of Computer, Hangzhou Dianzi University, Hangzhou 310018, Zhejiang, China)

Abstract In this paper it proposes a new improved fuzzy association rules algorithm by integrating Apriori and Kuok's algorithms. This algorithm, by introducing fuzzy membership function, decision tree scheme and similarities of rule sets, employed these three structures to build amended mining algorithm for mining association rules of quantitative attributes. Experimental results have demonstrated the algorithm's good performance in both rule building efficacy and time efficiency.

Keywords Intrusion detection Data mining Fuzzy logic Association rules

0 引 言

入侵检测系统(IDS)作为网络安全体系中的重要组成部分,通过检测并响应网络系统中的恶意行为,提高了系统的可靠性和生存性。

目前数据挖掘应用到入侵检测中时普遍采用的一种数据挖掘方法就是关联规则分析[1]。根据属性值的不同,关联规则可分为布尔型关联规则和数值属性关联规则。解决数值属性关联规则的一般方法是将数值属性关联规则挖掘问题转化为布尔型关联规则挖掘,但转化时存在布尔型关联规则问题(即边界划分过硬问题)[2]。为了解决这个难题,文献[3]引入模糊集合理论来达到软化边界的目的。

本文阐述了在入侵检测中应用模糊关联规则挖掘的方法,首先给出一个采用改进的模糊关联规则挖掘算法的入侵检测模型。改进的挖掘算法在传统关联规则算法 Apriori和 Kuok的基础上,引入模糊理论解决了数值属性转化为布尔属性时遇到的"尖锐边界问题",同时结合决策树[3]产生大项集和规则集,然后根据相似度的大小触发不同程度的响应。最后以网络流量分析为例,详细描述了在入侵检测中运用模糊关联算法发现攻击的过程。

1 模糊关联规则在入侵检测中的应用

根据入侵检测的基本模型[4],本文提出一种采用模糊关联规则算法的基于网络的异常入侵检测系统如图 1所示。

此模型的基本思想是:分别对采集的原始网络数据和当前主机的网络数据进行预处理,通过模糊关联规则挖掘,分别建立

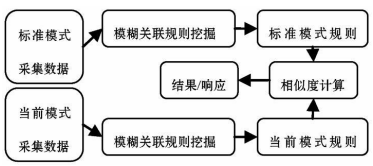

图 1 基于模糊关联规则挖掘的入侵检测模型

标准模式和当前模式下的规则集,然后计算两个规则集的相似度,根据不同的结果给出相应的响应。

1.1 模糊关联规则算法

Apriori[5]和 Kuok[6]算法最先应用到入侵检测领域,在此结合两者的优点,提出一种改进的算法:把数据划分为不同级别的集合然后生成一个关联规则集,算法使用大项集的概念增强规则的表述能力。

算法相关定义如下:

定义 1 $\{l_1, l_2, \ldots, l_n\}$是属性 a上的一个项目集,表示为 Terms(a),描述了 n条记录在属性 a上的一个值域集合。

定义 2 术语 t是一个序对$\langle a, l\rangle$这里的 $l\in$ Terms(a)(例如,$t=\langle a, l\rangle$那么 $t_a=a$并且 $t_l=l$表示 t在对应的属性 t_a上的值是 t_l)。

定义 3 一个项目集$\{t_1, t_2, \ldots, t_n\}$是这些项的一个有序集合,例如:如果 $\neq$ 则 $t_i\neq t_j$,如果 $<$ 那么 t_i应该排在 t_j之前。

用三个集合描述一个数据集:一个属性集 $A=\{a_1, a_2, \ldots, a_n\}$,对应的值域 $D=\{D_1, D_2, \ldots, D_n\}$和一个记录集 $R=\{r_1, r_2, \ldots, r_m\}$。例如:编号为 的记录 r_i在某个属性 a_j上的值域

收稿日期:2007-06-01。浙江省自然科学基金(Y106176);浙江省科技厅科技计划项目(2007C33058)。
费淑芳,硕士生,主研领域:信息安全。

表示为 $D_{ij}(1\leq i\leq m, 1\leq j\leq n)$。

定义 4 一个 i-项目集是当 $\{t_1, t_2, \ldots, t_i\}$ 的个数 $n=i$ 时的一个项目集。

定义 5 $x=\{t_1, t_2, \ldots, t_i\}$ 是一个 i-项目集，它的支持度 $\sup(x)$ 的计算如下：$\sup(x)=\dfrac{\sum_{k=1}^{m}\prod_{j=1}^{i} f_{tj}(d_{kj})}{m}$，$m$ 是这个数据集中记录的个数，i 是 x 中数据项的个数，d_{kj} 是记录 r_k 在属性 t_j 上的值。函数 $f_{tj}(d_{kj})$ 的计算有两种可能：如果 t_j 是一个模糊属性就使用隶属函数（见 1.2）；如果是一个绝对的或者是二进制的属性就使用下式：

$$f_{tj}(d_{kj})=\begin{cases}1 & \text{如果 } t_{il}=d_{ki}\\ 0 & \text{其它}\end{cases} \quad (1)$$

定义 6 如果一个项目集的支持度等于或大于最小支持度（minsup），称为大项集 L_i。

定义 7 L_i 是有 i 项的大项集。

定义 8 两个 i-项目集：$x=\{t_1, t_2, \ldots, t_i\}$ 和 $y=\{t_1, t_2, \ldots, t_i\}$，如果 $x.t_j=y.t_{j+1}$，$x.t_a\neq y.t_a$ 并且 $j=2, 3, \ldots, i$，那么它们可以结合为一个 $(i+1)$-项目集 $z=\{x.t_1, x.t_2, \ldots, x.t_i, y.t_i\}$，其中模糊属性对应的支持度为 $\sup(x.t_j)\wedge\sup(y.t_j)$。

定义 9 由 i-大项集 x 生成形如 $p\rightarrow q$ 的规则 r 的置信度 $\mathrm{con}(r)$ 用公式表示为 $\mathrm{con}(r)=\dfrac{\sup(x)}{\sup(p)}$，其中 $x=p\cup q$，$i\geq 2$。

定义 10 给定两条规则 R_1：$X\rightarrow Y(s, c)$ 和 R_2：$X'\rightarrow Y'(s', c')$，引入距离概念，$d(R_1, R_2)=\sqrt{(s-s')^2+(c-c')^2}$，定义这两条关联规则 R_1 和 R_2 的相似度为：

$$\mathrm{similarity}(R_1, R_2)=\begin{cases}0 & X\neq Y \text{ 或 } X'\neq Y'\\ 1-d(R_1, R_2) & X\neq Y \text{ 且 } X'\neq Y'\end{cases} \quad (2)$$

定义 11 两个规则集的 S_1 和 S_2 的相似度为：

$$\mathrm{similarity}(S_1, S_2)=\frac{s}{|S_1|}\cdot\frac{s}{|S_2|}=\frac{s^2}{|S_1||S_2|} \quad (3)$$

其中，$s=\sum_{\substack{\forall R_1\in S_1\\ \forall R_2\in S_2}}\mathrm{similarity}(R_1, R_2)$，$|S_1|$、$|S_2|$ 分别为 S_1 和 S_2 包含的规则数量。

改进的算法描述如下：

输入：最小支持度 minsup，最小置信度 minconf，记录集，隶属函数

输出：规则集

1）根据处理后的数据初始化树，设为 t；

2）求 1-大项集：

计算每一项的支持度，同一属性取较大的模糊区间，然后把支持度大于或等于 minsup 的项加入侯选 1-项目集，最后得到的是 1-大项集 L_1，初始化大项集 L 为 L_1；

3）根据 i-大项集 L_i 求 $i+1$-大项集 L_{i+1}

当 L_i 的长度大于 1 时，进行以下循环

{ 初始化 L_{i+1} 为空集；

检查属于 L_i 的任意两个项 x 和 y

{ 如果满足 $x.t_2=y.t_1$，$x.t_3=y.t_2\ldots$ 和 $x.t_i=y.t_{i-1}$，

则产生侯选 $(i+1)$-项目集 z；

如果 z 的支持度大于 minsup 则把 z 分别添加到 L_{i+1} 和 L；

}

加 1 后重复步骤 3，直到 L_{i+1} 为空或者 L 中项的个数等于全部属性的总数；

4）检查所有属于 L 的非空子集 x

{ 如果满足 p 和 q 是 x 的非空子集且 $q=x-p$

{ 如果规则 $p\rightarrow q$ 的置信度大于或等于 minconf，

则把这条规则加入规则集中，}

}

最后输出的是一个规则集。

1.2 构造隶属函数

我们仅仅只是为了验证方法的可行性，所以只选择了与网络流量相关的四个模糊属性来对系统进行分析：TCP和 UDP包在全部数据包中的比例 P_{TCP} 和 P_{UDP}，网络中每秒的平均数据包数量 Avg_packet/sec 以及每秒平均数据位 Avg_Mbit/sec。将四个数值属性划分为 High 和 Low 两个模糊集，然后在正常状态下，获取 100组网络数据，利用梯形分布的隶属函数形式[7]分别制定隶属函数。

1.3 实例分析

本节根据前面提出的模糊关联规则的改进算法进行实验，并在处理过程中给出程序的实例计算。利用 WinPcap软件，每一小时取一次样，共得到六组数据（如表 1所示为其中一组数据）作为标准模式下的训练数据集。下面以这一组数据为例描述模糊关联规则集的产生过程。

表 1 步骤 1的实验数据

编号	P_{TCP}(%)	P_{UDP}(%)	Avg_packet/sec	Avg_Mbit/sec
1	97.4	0.6	74.735	0.567
2	98.5	0.2	113.673	0.860
3	96.6	0.8	94.098	0.634
4	94.4	0.2	41.457	0.301
5	98.0	0.4	101.824	0.754
6	95.7	0.3	65.019	0.330
7	96.7	0.3	84.433	0.630
8	97.2	0.4	110.88	0.785
9	97.8	0.5	126.395	0.886
10	97.1	0.6	96.043	0.653

这是一个应用模糊关联规则算法进行数据挖掘的简单例子。根据定义 1、2和 3得到一个记录集 $R=\{r_1, r_2, \ldots, r_{10}\}$，$r_k=d_{k1}\ d_{k2}\ d_{k3}\ d_{k4}\ (1\leq k\leq 10)$。我们的目标是在 minsup=0.4 和 minconf=0.8的条件下发现关联规则。

步骤 1：初始化

为了表述方便，分别将 P_{TCP}、P_{UDP}、Avg_packet/sec 和 Avg_Mbit/sec 的模糊集记为：T.L、T.H、U.L、U.H、AP.L、AP.H、AM.L 和 AM.H。根据处理后的数据初始化建立一棵树：属性集的每个属性作为根的子节点，它的每个模糊集作为其子节点（如图 2所示）。

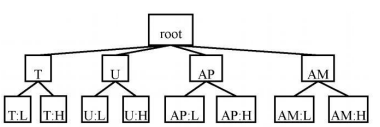

图 2 步骤 1的决策树

步骤 2：发现 1-大项集 （下转第 113页）

参 考 文 献

[1] 孟小峰. Web信息集成技术研究[J]. 计算机应用与软件, 2003, 20(11): 32-36.

[2] Juan Raposo, Alberto Pan, Manuel Alvarez, Justo Hidalgo. Automatically maintaining wrappers for semi-structured web sources[J]. Data & Knowledge Engineering, 2007, 61: 331-358.

[3] Knoblock C A, Lerman K, Minton S, Muslea I. Accurately and Reliably Extracting Data from the Web: A Machine Learning Approach. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2000, 23(4): 33-41.

[4] 刘迁, 焦慧, 贾惠波. 信息抽取技术的发展现状及构建方法的研究[J]. 计算机应用研究, 2007, 24(7): 6-9.

[5] 王敬普, 林亚平, 周顺先, 等. 基于包装器模型的文本信息抽取[J]. 计算机应用, 2007, 26(3): 655-658.

[6] 杜冬梅, 许彩欣, 苏健. 浅谈正则表达式在 web系统中的应用[J]. 计算机系统应用, 2007(8): 87-90.

[7] Steven Holzner. XML完全探索[M]. 北京: 中国青年出版社, 2001: 7-10.

[8] 邹晓静, 刘伟, 卢贤玲. 基于 XML的异构数据库系统的研究[J]. 微计算机信息, 2007, 23(8-3): 160-162.

[9] 贺令亚, 柳佳刚. 基于 Web的包装器技术的现状与发展[J]. 电脑开发与应用, 2007, 20(6): 27-29.

[10] 师雪霖, 牛振东, 宋瀚涛. 基于中介器/包装器的联合数字图书馆集成信息检索机制[J]. 计算机应用, 2005, 25(3): 703-705.

[11] 张素智, 李宏伟, 李树凯. 基于 Web服务的数据集成[J]. 郑州轻工业学院学报, 2005, 20(4): 34-37.

(上接第 96页)

根据隶属函数和定义 5计算每种属性对于模糊集的隶属度及其支持度, 得到新的数据集(见表 2)。

表 2　步骤 2的实验数据

T.L	T.H	U.L	U.H	AP.L	AP.H	AM.L	AM.H
10.357	10.643	10.143	10.857	11.0	20.806	10.446	10.554
20.0298	20.97	21.0	31.0	20.194	30.351	30.251	21.0
30.595	30.405	41.0	50.381	30.649	50.531	41.0	30.749
41.0	50.821	50.619	60.143	41.0	70.126	61.0	51.0
50.179	60.137	60.857	70.143	50.469	80.723	70.262	70.738
60.863	70.435	70.857	80.381	61.0	91.0	100.195	81.0
70.565	80.583	80.619	90.619	70.874	100.396		91.0
80.417	90.762	90.381	100.857	80.277			100.805
90.238	100.554	100.143		100.604			
100.446							
0.469	0.531	0.562	0.438	0.607	0.393	0.315	0.685

根据定义 6获取 1—大项集然后画出新的决策树(如图 3所示)。

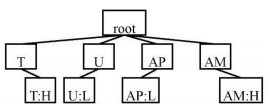

图 3　步骤 2的决策树

步骤 3　根据定义 8获取所有的大项集 L_n, $2 \leqslant n \leqslant 4$。

步骤 4　生成规则

所有大项集 $L=\{T.H, U.L, AP.L, AM.H, T.H, AM.H\}$, 根据定义 9产生规则, 最后得到一个规则集 S: $T.H \rightarrow AM.H$ (0.508, 0.957)。

最后获取网络正常和异常两种状态下的数据, 分别采用上述方法产生正常模式规则集 S_1和异常模式规则集 S_2, 然后根据定义 10和 11分别计算 S_1和 S_2与 S的相似度(见表 3)。

表 3　规则和规则集的相似度

规则集	规则	规则相似度	规则集相似度
S_1	$T.H \rightarrow AM.H$	0.9950	0.9900
S_2	$T.H \rightarrow AM.H$	0.8777	0.3852
	$AM.H \rightarrow T.H$	0	

最后根据相似度值的不同, 输出不同的结果:

(1) similarity $\geqslant$ 0.75时不响应;

(2) 0.5 $\leqslant$ similarity $\leqslant$ 0.7时提醒用户可能受到网络入侵;

(3) similarity $<$ 0.5时建议用户中断所有连接。

2 结束语

将模糊技术应用于入侵检测展现了人工智能技术在网络安全上的应用。本文结合了 Apriori和 Kuok的数据挖掘算法, 提出了改进的模糊关联规则挖掘算法。该算法解决了当前主流关联规则生成算法应用到入侵检测结果集的过程中所存在的数值属性区间划分过硬、扫描的数据集冗余和在第一阶段产生了众多与最后生成的规则集无关的频繁集等问题。

实验结果表明改进的模糊关联规则算法应用于入侵检测具有良好的效率。但是, 在实际的应用过程中我们还要做更多的工作。

参 考 文 献

[1] Jiawei Han, Micheline Kamber. 范明, 孟小峰, 译. 数据挖掘概念与技术[M]. 北京: 机械工业出版社, 2003.

[2] Rakesh Agrawal, Tomasz Imielinski, Arum Swami. Mining association rules between sets of items in large databases[C]. Proc of the ACM SIGMOD Conference on Management of Data. Washington, D.C., 1993.

[3] Aly El-Semary, Janica Edmonds, Jesus Gonzalez-Pino, et al. Applying data mining of fuzzy association rules to network intrusion detection: proceedings of the 2006 IEEE Workshop on Information Assurance[C]. United States Military Academy, West Point, NY, 2006.

[4] 徐蕾, 刘宝旭. 基于数据挖掘技术的入侵检测系统设计与实现[J]. 计算机工程, 2002, 28(6): 9-11.

[5] Agrawal R, Srikant R. Fast algorithms for mining association rules: 20th International Conference on Very Large Databases[C]. Santiago, Chile, 1994.

[6] Kuok C, Fu A, Wong M. Mining fuzzy association rules in databases[S.l.], 1998.

[7] 王季方, 卢正鼎. 模糊控制中隶属函数的确定方法[J]. 河南科学, 2000, 18-4.