一种基于关联规则挖掘的入侵检测系统

王文槿[1,2],刘宝旭[1]

(1. 中国科学院高能物理研究所计算中心,北京 100049; 2. 中国科学院大学,北京 100049)

摘要: 为解决入侵检测系统普遍存在漏/误报率高、特征库需频繁升级、更新等问题,根据中科院高能所的网络环境,构建了一种基于关联规则挖掘算法的入侵检测系统。该系统可以通过训练数据生成通用的检测规则,并利用规则检测新的攻击。实验证明:该系统对未知攻击具有较好的检测速度和检测率。

关键词: 入侵检测; 关联规则挖掘; Apriori; K-Means 聚类; FP-Growth

中图分类号: TP 393.08 **文献标志码:** A **文章编号:** 0258-0934(2015)02-0119-05

近年来,随着我国高能物理的研究的发展逐渐引起世界关注,围绕中科院高能所科研网络的恶意攻击行为日渐增多(如:恶意代码、蠕虫、拒绝服务、钓鱼网站等),大量保密实验数据的安全问题日益严重。为此,构建一套行之有效的网络入侵检测系统(IDS)已成为科研单位网络信息系统的当务之急。

IDS代表的是一套能够在系统被攻击时发出警告的防卫工具的统称[1]。传统IDS通常采用模式匹配、统计分析、完整性分析等方法,对网络行为和系统进行实时的监控和检测。模式匹配因为其准确性高,速度快等优点,对于检测已知的攻击非常有效,但是无法检测未知类型的攻击。统计方法一般都是将用户的正常状态和合法行为建模,然后将用户活动和正常模式比较,判断入侵的行为。统计方法虽然可以一定程序弥补不能检测未知攻击等缺点,但误报率和漏报率都非常高[2]。

收稿日期:2014-03-05

基金项目:国家科技项目(2012BAH14B02)和国家信息安全专项(发改办高技[2012]1424号)资助。

作者简介:王文槿(1988-),男,云南昌宁县(市)人,硕士研究生,主要从事网络安全研究。

数据挖掘技术可以从海量数据中发现有意义的知识,将数据挖掘技术应用于网络审计数据的分析,生成的有意义的知识可以构建更好的入侵检测模型[3]。本项目基于关联规则挖掘,提出了新的Apriori算法和FP-Growth算法,实现了规则的提取和海量数据的实时检测方面都有较好的效率,获得了对未知类型的攻击检测有较高的正确率。

1 关联规则挖掘的基本原理

1.1 基本原理和定义

关联规则挖掘可发现大量数据项集之间有趣的关联或相关关系。关联规则的获取是通过某个数据挖掘方法从事务数据库中找出隐含的频繁模式。在入侵检测技术中就是从海量的网络流量数据中挖掘出相应的频繁模式来判断当前网络是否正常。关联规则的挖掘主要有两种方法:Apriori和Fp-growth。与关联规则挖掘有关的两个重要概念就是项集的支持度和置信度[4],定义如下:

(1) 项集的支持度:数据库中包含(支持)项集X的事务的数目称为项集X的支持度计数,记为$\sigma(X)$。项集X的支持度定义为式(1),其中T为事务的集合。

$$\sigma(X) = |\{t_i \mid X \subseteq t_i, t_i \in T\}| \quad (1)$$

(2) 规则的置信度: 关联规则 X - > Y 的置信度代表 Y 在包含 X 的事务中出现的频繁程度,可由式(2)计算得到:

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \quad (2)$$

支持度和置信度是描述关联规则的两个重要概念,前者用于衡量关联规则在整个数据集中的统计重要性,后者用于衡量关联规则的可信程度。一般来说,只有支持度和置信度均较高的关联规则才可能是用户感兴趣、有用的关联规则。

1.2 新型 Apriori 算法描述

Apriori 算法是第一个关联规则挖掘算法,由 Agrawal 提出[5]。它开创性的使用基于支持度的剪枝技术,系统地控制候选项集的指数增长。算法利用第 k 次循环产生的最大长度为 k 的频繁项集构造长度为 k + 1 项的频繁候选集,然后扫描数据库,计算候选项集的支持数,得到 k + 1 项频繁项集,直到某次循环最大项集为空时算法停止。

候选项集的生成是 Apriori 算法的核心,它基于向下闭包的性质,即如果一个项集是频繁项集,那么它的所有子集也是频繁项集。反之,如果一个项集的某个子集不是频繁项集,那么这个项集也一定不是频繁项集。因此,由 k 项频繁集构造 k + 1 项频繁候选集时包括连接和剪枝两个步骤。

该算法的不足是频繁项集的长度每增加一个,都要遍历一次数据库。且随着频繁项集长度的增加,候选项集数目逐渐减少,包含这些候选项集的事务数据量也越来越少,但在 Apriori 算法中并没有减少事务数据库的扫描范围[6]。

考虑到传统 Apriori 算法在候选集生成上的不足,本项目提出了一种基于经典 Apriori 算法的新型 Apriori 算法,大大提高了算法在 IO 上的效率。该新型 Apriori 算法首先将事务编码成精简的字符串并全部读入内存中,然后从中生成由项集作为 key,对应的支持度计数为 value 的字典。然后在此基础上迭代生成所有频繁项集。

1.3 FP - Growth 算法描述

Apriori 算法在产生频繁模式完全集前需要对数据库进行多次扫描,同时产生大量的候选频繁项集,即便按照 1.2 中所描述方法进行适当优化,但总体上该算法的时间和空间复杂度都比较大。但是 Apriori 算法中有一个很重要的性质: 频繁项目的所有非空子集也必须是频繁的。基于此性质,Jiawei Han 提出了 FP - Growth 算法。

FP - Growth 算法的核心是构造 FP - Tree,FP - Tree 是频繁模式树,它将整个事务数据库压缩到一棵频繁模式树上。而且,在构造事务数据库的 FP - Tree 的过程中,只需要扫描一次事务数据库就能生成,比 Apriori 算法节省很多的时间。FP - Growth 算法有两个步骤,首先是构造 FP - Tree,然后在该 FP Tree 上挖掘出频繁项集。

2 基于关联规则挖掘的入侵检测技术

关联规则挖掘算法处理的对象一般都是由二元属性组成事务数据库,但是网络流量统计数据则大多都是连续的数据。表 1 所示的是一个常规流量统计数据信息,可以看到,除了 service 和 flag 属性是离散的值外,其他的如 time,duration,src_bytes 等属性都是连续的值。在这样的情况下,如何正确科学地将网络流量统计数据离散化就成为将数据挖掘方法应用于入侵检测系统的一个主要挑战。

表 1 常规的流量统计信息

time	duration	service	src_bytes	dst_bytes	flag
1.1	10	telnet	100	2000	SF
2.0	2	ftp	200	300	SF
2.3	1	smtp	250	300	SF
3.4	60	telnet	200	12100	SF
3.7	1	smtp	200	300	SF
3.8	1	smtp	200	300	SF
5.2	1	http	200	0	REJ
3.7	2	smtp	300	200	SF

2.1 数据预处理

将连续数据离散化,最常见的方法就是直接将数据按照一定的标准分组,形成有限个区间。分组的标准一般由用户根据数据的特点手工指定。但由于网络流量信息一般都是海量的数据,所以需要使用更加智能的自动化离散方法。

本项目采用了 K - means 聚类方法,通过数据分布等相关信息来将连续的流量数据划分到不同的离散类别中去。该方法的主要作用是

将输入数据集合按照同一个类别相似度较高，不同的类别相似度较低的标准划分为 k 个类别。相似度的定义为对每个类别的对象求其均值，然后将均值作为该类型的中心，进而计算出该类别的其他对象的相似度[7]。

K－means 算法工作的基本流程如下：(1) 从输入的数据集中任意选择 k 个作为中心；(2) 计算所有对象与上述 k 个对象的相似度之后，依次排序将每个对象都归类到相似度最小的类别中去；(3) 对所有类别重新计算平均值；(4) 重复 step2－3 直到均方差测度函数收敛。

本项目用 k－means 方法处理所有网络流量统计数据中的连续属性，将其分别映射到有限个类别中，并对每个类别进行编码。通过这样的方式将属性较为复杂的流量数据编码成一个个符号组成的字符串。

例如：假设离散化系数为 3，且仅考虑 src_bytes 和 dst_bytes 两个属性，若 src_bytes 的聚类结果的质心为 (480.052226，693375640，66407.895091)，dst_bytes 的聚类结果质心为 (5074279.3125，2217.760389，1509908.315789)，上述六组质心对应六个类别，分别用 a～f 来表示。若一个网络连接的数据为：src_bytes = 450，dst_bytes = 1600000. 则相应编码为 af。

2.2 系统设计

在将连续的属性数据离散化之后，可以通过 Apriori/FP Growth 算法来挖掘海量网络流量数据中的频繁模式，并生成可以被用来检测未知攻击类型的强规则。但在传统关联规则挖掘算法的思想中，只有置信度高的规则才会被认为是值得关注的，这样的思想并不适用于入侵检测系统。因为一个攻击不管多小都有可能造成很大的危害，所以在规则的产生中不能忽视一些相对置信度不高的攻击规则。要解决这个问题，比较科学的方法就是首先通过适当降低“强规则”的门槛，使用较低的置信度生成大量的规则，然后再根据规则所对应的攻击类型对结果集进行剪枝，最后得到一个较为全面且不冗余的规则库。

综上所述，可以得出基于关联规则挖掘算法的入侵检测系统流程如图 1 所示。系统处理海量的历史网络审计数据，对于每一条连接信息，使用 k－means 算法将其对应的连续值离散化进而将整条连接信息编码为字符串表示。接下来使用 Apriori/FP Growth 算法对之前处理的字符串结果进行挖掘，挖掘的规则都输出到规则库中。

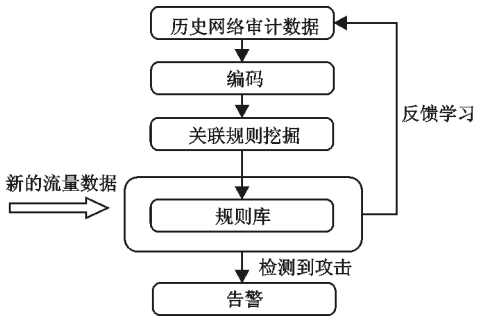

图 1 基于关联规则挖掘算法的入侵检测系统流程图

规则库同时也是实时检测模块的一部分。实时检测模块不断地监听网络的流量数据，对于每一条连接信息都将其与规则库中的检测规则进行比对，如果发现攻击则触发告警，若没有发现则将当前连接信息存入历史网络审计数据参与下一轮的规则挖掘。

3 实验设计

实验采用 C#语言，基于 VisualStudio2010 实现了 Apriori 和 FP Growth 算法，并对两个算法的性能进行了比较。基于此实现了一个入侵检测系统，并采用 KDDCup99 的数据集对该系统进行了测试。

3.1 KDDCUP99 数据集

KDDCUP 是 1999 年 KDDCUP 知识发现竞赛的数据集，由美国 DARPA 在 MIT 林肯实验室的模拟实验数据集修改而来。在 KDDCUP99 的数据集中，一个网络连接定义为在某个时间内从开始到结束的 TCP 数据包序列，并且在这段时间内，数据在预定义的协议下(如 TCP、UDP) 从源 IP 地址到目的 IP 地址的传递。每个网络连接被标记为正常(normal) 或异常(attack)，异常类型被细分为 4 大类共 39 种攻击类型，其中 22 种攻击类型出现在训练集中，另有 17 种未知攻击类型出现在测试集中。

4 种异常类型分别是[8]：(1) DOS，denial－of－service 拒绝服务攻击，例如 ping－of－death，syn flood 和 smurf 等；(2) R2L，unauthorized access from a remote machine to a local machine 来自远程主机的未授权访问，例如 guessing password；(3) U2R，unauthorized access to

local superuser privileges by a local unpivileged user 未授权的本地超级用户特权访问，例如 buffer overflow attacks；(4) PROBING，surveillance and probing，端口监视或扫描，例如 port – scan，ping – sweep 等。

虽然年代有些久远，但 KDD99 数据集仍然是网络入侵检测领域的事实 Benchmark，为基于计算智能的网络入侵检测研究奠定基础。前 10^5 条记录中攻击类型的分布如表 2 所示。

表 2　KDDCUP 前 10^5 条记录攻击类型分布情况

项目	次数	项目	次数
normal	56237	ipsweep	760
buffer_overflow	5	land	17
loadmodule	2	ftp_write	8
perl	2	back	2002
neptune	20482	imap	12
smurf	19104	satan	539
guess_passwd	53	phf	3
pod	40	nmap	231
teardrop	199	multihop	6
portsweep	278	warezmaster	20

3.2　算法实现

3.2.1 聚类算法的实现

本课题采用 Visual C + + 实现了 K – means 算法，可以对 3.1 中描述的 KDDCUP 数据集合中的连续数据进行离散化，程序的核心结构如下所示：

```
//拷贝质心数组到副本
void CopyCenter( )
//初始化质心，随机生成法
void InitCenter( )
//加入一个数据到一个 Cluster [index] 集合
void AddToCluster( int index,double value)
//重新计算簇集合
void UpdateCluster( )
//重新计算质心集合，对每一簇集合中的元素加总求平均即可
void UpdateCenter( )
//判断 2 数组元素是否相等
int IsEqual( double *  center1 ,double *  center2)
//打印聚合结果
void Print( )
//初始化聚类的各种数据
void InitData( )
```

3.2.2 关联规则挖掘算法的实现

为了方便演示关联规则挖掘算法的过程，实验使用 Visual C#制作了可视化的版本，通过该系统可以实现关联规则的生成，以及基于规则对入侵数据进行入侵检测。如图 2 所示。

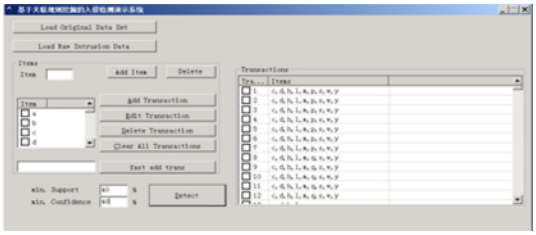

图 2　演示系统运行效果图

系统可以载入已经离散化之后数据，并转化成为会话集合。左侧的功能区域可以对会话集合进行常见的添加，删除，修改等功能。

3.3　结果分析

实验从 kddcup. data_10_percent_corrected 数据集中提取了 100000 条记录，分别用 Apriori 算法和 FP Growth 算法生成相应的入侵检测规则，结果如表 3 所示。然后基于提取出的规则对未标记的测试数据进行了不同量级的测试，结果如表 4 所示。

表 3　Apriori 算法和 FP Growth 算法结果

算法	生成检测规则时间/s
Apriori 算法	18.8448000
FP Growth 算法	2.74603678698

表 4　未知攻击检测算法结果

测试数据范围	检测正确率
10000 条	92.62%
100000 条	87.691%
200000 条	86.7025%
300000 条	87.1486%

从表 3 中可以看出：FP Growth 算法拥有 6 倍于 Apriori 算法的速度，所以在实际的入侵检测系统中推荐使用 FP Growth 算法。如表 4 所示，在前 10000 条数据中，本算法可以达到 92% 的准确率。在之后的 100000 条，200000 条和 300000 条数据的测试中，本算法一直保持了高达 87% 的准确率，较好地解决了传统算法不能检测未知攻击，以及统计分析不能保证检测的正确率等问题，其优势得以明显体现。

4　结论

本项目基于关联规则挖掘算法构建了一种新型入侵检测系统。该系统通过 k – means 聚类方法对连续数据进行离散化，然后利用关联

规则挖掘算法，实现对未知和已知攻击入侵的检测。实验在对测试数据进行测试的基础上，对具体算法的选择也进行了性能的比较。实验结果表明：优化后的算法比传统的入侵检测系统更能准确地检测未知和已知攻击，同时具有较快的检测速度。不过，如何更加精确地检测出攻击的类型和根据攻击类型采取适当的应急措施，仍是进一步研究的方向。

参考文献：

[1] Flora S，Tsai. Network Intrusion Detection Using Association Rules [J]. International Journal of Recent Trends in Engineering，2009，l 2(2)：202 - 203.

[2] 胡昌振. 网络入侵检测原理与技术 [M]. 北京：北京理工大学出版社，1996.

[3] 陈树刚. 关联规则挖掘在入侵检测系统中的研究 [D]. 哈尔滨工程大学硕士论文，2007.

[4] 王怡，谢俊元. 入侵检测系统中关联规则挖掘技术的研究 [J]. 计算机科学，2008，l35(10)：81 - 82.

[5] Agrawal R，Imielinski T，Swami A. Mining association rules between sets of items in large databases [C]. ACM SIGMOD Conference on Management of Data (SIGMOD93)，1993，4：307 - 328.

[6] 王宗晨. 基于数据挖掘的日志审计系统研究与实现 [D]. 清华大学硕士论文，2008.

[7] Pang - Ning Tan，Michael Steinbach，Vipin Kumar. Introduction to Data Mining [M]. 北京：人民邮电出版社，2011.

[8] Mahbod Tavallae，Ebrahim Bagheri，Wei Lu and Ali A. Ghorbani. A Detailed Analysis of the KDD CUP 99 Dataset [C]. Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009).

Association Rule - Based Network Intrusion Detection System

WANG Wen - jin[1,2]，LIU Bao - xu[1]

(1. Computing Center，Institute of High Energy Physics，Beijing 100049，China；
2. University of Chinese Academy of Sciences，Beijing 100049，China)

Abstract: In order to fix the problems existed in traditional IDS，e. g. high leak rate detection/false alarm rates and feature library needs frequently upgrade，based on IHEP network environment，this paper presented a new Intrusion Detection System based on association - rule mining algorithm. This system can detect unknown attack by generating rules from training data. Experiment proved that the system has great accuracy and performance on both unknown and known attack detection.

Key words: intrusion detection；association rule mining；Apriori；K - means clustering；FP - growth