TCP/IP 审计数据缩减技术在入侵检测中的可行性研究

田俊峰 王惠然 傅 玥
(河北大学网络技术研究所 保定 071002)

摘 要：目前的一些入侵检测系统是利用网络层的 TCP/IP 数据包里的特征进行分析建模，但 TCP/IP 的特征属性对检测过程的贡献不同，因而如果能够在不影响检测准确性的前提下，适当缩减特征属性的数量，那么对于提高 IDS 的检测率和实时性势必产生有益的影响，鉴于此该文提出基于决策树的规则统计方法(DTRS)来缩减 TCP/IP 的特征属性。它的基本思想是通过在 n 个子数据集上建立 n 棵决策树，提取其中的规则，根据特征属性使用频度的不同，计算出相对重要的属性，并通过实验验证了其可行性和有效性。

关键词：入侵检测；特征缩减；决策树
中图分类号：TP393.08 **文献标识码**：A **文章编号**：1009-5896(2007)09-2248-04

Research on the Feasibility of TCP/IP Feature Reduction for Intrusion Detection

Tian Jun-feng Wang Hui-ran Fu Yue
(*Institute of Network Technology, Hebei University, Baoding* 071002, *China*)

Abstract: At present some Intrusion Detection Systems (IDS) use the features of TCP/IP data packets for analysis and modeling, but due to the different contribution of TCP/IP features to the detecting process a favorable impact may be made on the promotion of IDS's detecting rate and real time if the quantity of properties can be reduced properly without affecting the precision of detection. Therefore, a Decision Tree Rule-based Statistical method (DTRS) in light of this is presented to reduce TCP/IP features. Its primary concept is to create n decision trees in n data subsets, extract the rules, work out the relatively important features in accordance with the frequency of use of different features and verify its feasibility and effectiveness through tests.

Key words: Intrusion detection; Feature reduction; Decision tree

1 引言

随着Internet的快速发展，除了传统的防火墙、加密认证系统等入侵防御方法之外，入侵检测系统(Intrusion Detection System， IDS)作为抵御网络入侵攻击的重要手段已经得到广泛的研究和应用[1]。入侵检测技术指的是用来检测违反安全规范的一种机制[2]。入侵检测系统大致可分为两类：基于主机的入侵检测和基于网络的入侵检测。基于主机的入侵检测主要是使用日志跟踪，而基于网络的入侵检测主要是通过对网络上数据包的分析来进行。基于网络的入侵检测对连接中数据包的信息进行分析，提取数据包中的特征信息，分析攻击行为的特征，再根据一定的规则来创建系统。在过去的二十年里，入侵检测技术取得了很大的发展，但仍远远不能满足计算机系统安全需要。检测率偏低、误报率偏高仍然是入侵检测系统的主要问题，即便准确性稍高的检测系统仍存在着时间和资源浪费的问题[3]。很多入侵检测系统都努力尝试在实时环境下来完成检测任务，但是由于庞大的审计数据，使得计算量非常大，准确率降低，从而很难实现实时检测，因此鉴别并选取重要的输入特征对于IDS意义重大。

文献[4]中，Mukkamala 和 Sung 利用支持向量机技术效能等级排序算法(PFRM)做了对 TCP/IP 数据包中重要特征的鉴别选取，并用支持向量机(SVM)和神经网络(ANN)分别对缩减的审计数据进行实验比较，得出了下面的结论：只使用部分特征比缩减前提高了检测率，缩短了训练和测试时间；并且 SVM 在检测率和测试时间上都明显优于 ANN。但是 PFRM 在对某个特征做重要性测试时，只考虑了该特征被去掉后的性能，未考虑特征之间的相互联系。文献[5]中，Chebrolu，Abraham 和 Thomas 利用贝叶斯学习和马尔可夫毯(BLMB)模型的方法，以及回归树学习的方法(CART)分别做了对重要特征的鉴别选取，并进行了测试分析，BLMB 和 CART 对于 4 种攻击类型(Dos,Probe,U2r,R2l)的测试精度不同，准确率相差很大，特别是 U2r 的测试精度明显低于 PFRM。文献[6]中，邹涛等利用基于遗传算法的最小特征子集选取算法(FSSGA)，在特征减少和攻击检测方面的性能比 PFRM 有所提高，但总的检测率偏低。

本文针对重要特征的鉴别选取问题，提出了基于决策树的规则统计的方法(Decision Tree Rule-based Statistic，

2005-12-16 收到，2007-05-21 改回
河北省自然科学基金(F2004000133)资助课题

DTRS)，采用 MIT 的 Lincoln 实验室的审计数据进行实验，对测试结果进行了分析，在数据缩减和分类精度上取得了令人满意的结果。

2　决策树规则统计法(DTRS)及测试方法

2.1　DTRS 基本思想

决策树(Decision Tree，DT)是数据挖掘(Data Mining，DM)的一种方法，通常是指从海量的数据中挖掘出有用的模型[7]，因此对于一些重要的信息，有集中体现的作用。决策树是一种逼近离散值目标函数的方法，在这种方法中学习到的函数被表示为一棵决策树，学习得到的决策树也可以被表示为多个if-then的规则[7]。在以往用决策树做实验过程中，发现这样一个特性，决策树生成的规则，对属性的使用频度相差很大，有的几乎每条规则都用到，有的在规则中没有用到或用到的次数非常少，而且对于不同的类别，所使用的判定属性也不一样，鉴于此，根据属性的使用频度，只提取出在规则判定中相对重要的属性，这就是下面将要谈到的决策树规则统计方法(DTRS)。

2.2　DTRS 实现方法

(1) DTRS 算法理论分析　决策树可以将分类信息用 if-then 规则的形式来表示，而这些规则本身即为知识的一种表现形式，即为知识的载体；决策树的生成过程中，要对属性的重要性(即该属性对于分类的贡献大小)进行选取，以确保建立的树有很好的分类能力。DTRS 算法基于决策树的信息增益方法，是在信息增益算法之后的进一步的判别，因此，对于鉴别重要属性是有一定价值的。DTRS 方法是将海量数据分成若干个数据子集，在每个数据子集上建立一棵决策树，提取出其中的规则集，对规则中出现的特征属性按类别进行统计，最后把不同的规则集上统计出的特征属性，按类别进行合并。

(2)DTRS 算法　DTRS 算法的流程图如图 1 所示。

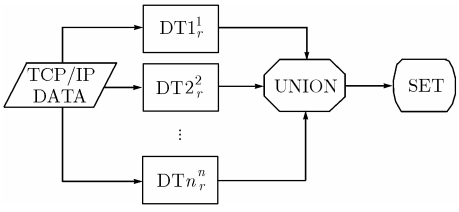

图 1 DTRS 流程图

DTRS 算法步骤如下：

步骤 1　把海量的 TCP/IP 数据分成 n 个数据子集，根据 n 个数据子集建立 n 棵决策树(DT1,DT2,…,DTn)。

步骤 2　对于每一棵决策树，提取它的规则集 $(r^1,r^2,\cdots,r^n)$，对每一个规则集 r^i，根据类别对特征属性进行统计，再将 n 个规则集的结果分类取合。假设特征属性共有 $m(m[1],m[2],\cdots,m[m])$个，类别 class 有 $k(k1,k2,\cdots,kk)$类，$kk[j]$是第 kk 类对属性 $m[j]$设置的计数器，核心算法如下：

```
input:  (r^1,r^2,…,r^n);
output: k1[j],k2[j],…, kk[j];
while(ri 不为空 and i 小于等于 n)      //i 为 1 到 n 中的某一个值;
{
 for(count=firstRule; count<=lastRule; count++)      //每一个 ri 中有若干条 rule,count 是设置的循环变量;
 {
  switch(class)      //class 为类别,查看 rule 中的 class 为哪一类别;
    {
    case k1:for(j=1; j<=m; j++)      //查看所有属性是否该 rule 中出现;
            {
            If (m[j]不为空) k1[j]加 1;      //若该属性在 rule 中存在,相对应的计数器加 1;
             }
             break;      //跳出 switch 语句,继续 for 循环;
     ...      //省略其中 k2 到 kk-1 语句;
    case kk:for(j=1; j<=m; j++)      //查看所有属性是否该 rule 中出现;
             {
             if(m[j]不为空)kk[j]加 1;      //若该属性在 rule 中存在,相对应的计数器加 1;
              }
              break;
    }
}
i++;      //i 加 1, 继续 while 循环。
}
```

步骤 3　对每一类别$(k1,k2,\cdots,kk)$，根据上面统计的结果对 m 个属性进行排序，按其使用的次数的多少，从高到低排序。最后再针对所有的类别，把特征属性按其使用次数从高到低排序，排在后面的使用次数为 0 或者小于 2(经验值)的属性就可以忽略掉，只保留前面的 i 个属性($i<=m$)，这 i 个属性便组成了重要特征集 SET。

下面将用实验来证明，用 i 个属性的检测性能优于 m 个属性的检测性能。

2.3　测试方法

测试方法仍是用的决策树建立的分类器，具体方法如下：

(1)将数据集分成训练和测试两部分，训练子集由决策树生成 n 条规则，假设其中的一条规则为(A>3,B<=7,C=0 − >class G [0.833])，如果一条测试数据的特征同时满足

(A>3,B<=7,C=0)，那么这条数据的类别为 G 的可信度为0.833，根据 n 条规则用测试子集去测试，通过对多条规则的可信度作比较，类别的最后判定取决于可信度的最大值所属的类别。

(2)将训练子集分成 i 个特征属性和 m 个特征属性的子集，再生成不同的规则集，由测试子集进行测试，通过对两组测试结果的比较，可以验证 i 个属性的检测性能优于 m 个属性的检测性能。

3 实验仿真及结果分析

3.1 实验数据

实验数据采用的是MIT的Lincoln实验室的Tcpdump[7]数据，它是由DARPA(高级国防部高级计划研究署)收集，其目的是调查、评估现有入侵检测技术的性能。Lincoln实验室构造了一个局域网，模拟一个典型的空军LAN，以获得原始的TCP/IP数据，数据中包含了军事网络环境中的几乎所有攻击类型。Tcpdump数据每一条TCP/IP连接有 41 个特征属性，数据集中的攻击类型分为 4 大类：Dos(拒绝服务攻击)，Probe(隐秘探测)，U2r(对于本地的高级用户进行无授权的访问)，R2l(来自于远程的无授权访问)。这 4 大类又包含一些小的攻击类型，如表 1 所示。实验中，采用Tcpdump10%数据集中的一部分数据作为训练数据，一部作为测试数据。

表 1 10%数据集攻击类型分类表

Dos	Probe	U2r	R2l
Back	Ipsweep	Loadmodule	Multihop
Land	Nmap	Perl	Ftp_write
Smurf	Satan	Rootkit	Imap
Teardrop	Portsweep	Buffer_overflow	Phf
Neptune			Guess_passwd
Pod			Warezmaster

3.2 特征统计(DTRS)

DARPA 实验数据集中的 10%组成了一个单独的数据子集，把这个数据子集平均再分成 3 个小的数据子集，采用 See5 软件，建立了 3 棵决策树，再根据 3 棵决策树产生的规则，在 Delphi 平台下，按 DTRS 算法，得出如下的统计属性列表。表 2 是对各个特征属性在所有规则里面使用的次数统计，表 3 是根据表 2 统计的结果对各个特征属性按其使用的频率从高到低进行的排序，特征属性名用相对应的序号来代替。表 3 中，“总”的一项指的是对于 Normal 和 Attack 中所有类型，特征属性使用频率的排序，根据这一项，去掉使用次数为 0 和小于 2 的属性，得到最优特征子集(E,C, J, AH, AJ,F, AI, AG, AK, M, A, B, V, AM, AN, AO, W, AA, AC, AF, L, X, K, D)共 24 个属性，比原来减少了 41.46%，下面通过实验来证明 24 个特征属性的检测性能优于原来 41 个的检测性能。

表 2 特征属性使用频率统计表

特征属性	序号	U2r	R2l	Probe	Normal	Dos	总
duration	A	1	6	3	6	0	16
protocol_type	B	0	0	10	4	0	14
service	C	17	31	20	8	3	79
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
dst_host_diff_srv_rate	AI	0	13	8	3	2	26
dst_host_same_src_port_rate	AJ	6	20	16	13	1	56
dst_host_srv_diff_host_rate	AK	0	8	7	4	0	19
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
dst_host_srv_serror_rate	AM	4	2	6	1	0	13
dst_host_rerror_rate	AN	2	3	4	3	0	12
dst_host_srv_rerror_rate	AO	0	4	6	1	0	11

表 3 特征属性次序表

类别	特征属性的排序
Normal	AH,E,J,AJ,C,F,A,W,B,M,AC,AK,AF,AG,AI,AN,AA,AM,AO.
Dos	AH,E,J,C,AI,AJ
Probe	E,C,AJ,J,F,AH,B,AG,AI,AK,V,AM,AO,M,AC,AN,AK,W,AF,AA,G,H,L,P,X,AD,AE
U2r	C,J,E,AJ,AM,M,F,AH,AG,AF,AC,AN,AA,V,A,W,L,D
R2l	E,C,J,AH,AJ,AI,AG,AK,F,A,L,M,V,AA,AO,X,AC,AN,W,AM,AF,D,AE
总	E,C,J,AH,AJ,F,AI,AG,AK,M,A,B,V,AM,AN,AO,W,AA,AC,AF,L,X,K,D,AE,G,H,P,AD

3.3 实验测试过程及结果

(1)数据准备　训练数据随机选取 11590(Normal: 4163, Dos: 4574, Probe:1627, U2r:30, R2l:196)条进行训练实验，测试数据随机选取 3854(Normal:1781, Dos:1076, Probe:651, U2r:30, R2l:316)条进行测试。

(2)采用 See5 软件生成决策树并抽取规则

(a)按 DTRS 特征统计得到的最优特征子集(E,C, J, AH, AJ, F, AI, AG, AK, M, A, B, V, AM, AN, AO, W, AA, AC, AF, L, X, K, D)，共 24 个特征属性，把训练数据根据 24 个特征属性进行整理，并把类别整理成五类(Normal,Dos, Probe,U2r,R2l)，再使用 See5 软件生成一棵决策树，提取出其中的规则，记为规则 1。

(b)按BLMB算法得到的最优特征子集(A,B,C, E,G, H, K, L, N, Q, V, W, X, Y, Z, AD, AF)[5]，共 17 个特征属性，把同样的训练数据按上面同样的方法进行整理，生成决策树，提取规则，记为规则 2。

(c)最后使用未删减的特征属性(A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, AO)，共 41 个，把同一批数据按上面的方法进行整理，生成决策树，提取规则，记为规则 3。

(3)采用 Delphi 编程软件作为测试实验平台，用训练数据生成的规则来对每一条 TCP 数据测试，如规则 1 中的一条规则(A<= 3, E> 353, J<= 0, L= 1, – > class normal [0.999])，意思是如果一条 TCP 数据的特征同时满足 duraion <= 3，src_bytes> 353，hot <= 0，logged_in = 1，那么这条数据的类别为 Normal 的可信度为 0.999，通过对多条规则的可信度作比较，类别的最后判定取决于可信度的最大值所属的类别，由于 Dos，Probe，U2r，R2l 都是攻击类型，所以最后将这 4 种类型融合为 Attack 类型，这样忽略了 Dos，Probe，U2r，R2l 之间的差别，可以提高 Attack 的检测率。根据上面得到的规则 1，规则 2，规则 3，进行测试，得到下面的结果。

(4)实验结果与分析　表 4 显示的是使用不同特征子集建立的分类器对 TCP 数据集进行检测的分类准确率和错误率。

实验中准确率定义为攻击类型和正常类型各自被正确检测的概率，总检测率定义为所有类型被正确检测的概率，虚警率定义为正常数据错判为攻击类型的错误率。通过表 4 测试结果综合比较，可以看出 DTRS 在数据缩减、节省时间和分类精度上都取得了比较满意的结果。

4　结束语

DTRS 方法在缩减 TCP/IP 的特征属性方面，利用了数据挖掘能够自动考虑特征属性之间的相互联系，根据大量数据生成的规则，将集中的属性提取出来，再使用相对重要的特征属性构建入侵检测模型，不但节省了检测的时间，也提高了检测率，如果应用到实时系统中，将会提高系统的效率。不足的是，目前的研究仍是实验阶段，对于如何应用到实时系统，有待在以后的时间进行深入研究验证。

表 4 测试结果 (%)

测试标准	DTRS (24 个)	BLMB(17 个)	41 个特征属性
Normal 准确率	98.37	97.87	97.14
Dos 准确率	97.12	96.84	95.91
Probe 准确率	96.01	82.18	99.84
U2r 准确率	86.67	63.33	90.00
R2l 准确率	77.85	80.38	80.06
Attack 准确率	93.68	89.24	94.64
总检测率	95.85	93.23	95.79
虚警率	0.75	0.99	1.32
测试时间	8.39(s)	6.91(s)	12.58(s)

参 考 文 献

[1] 郑军，胡铭曾，云晓春，张宏丽. 基于 SOFM 和快速最近邻搜索的网络入侵检测系统与攻击分析. 计算机研究与发展，2005-9, 42(9): 1578-1586.
Zheng Jun, Hu Ming-zeng, Yun Xiao-chun, and Zhang Hong-li. Network intrusion detection and attack analysis based on SOFM with fast nearest-neighbor search. *Computer Research and Development*, 2005, 42(9): 1578-1586.

[2] Bierman E, Cloete E, and Venter L M. A comparison of intrusion detection systems[J]. *Computers & Security*, 2001, 20(8): 676-683.

[3] Lee W, Miller M, Stolfo S, Jallad K, Park C, Zadok E, and Prabhakar V. Toward cost-sensitive modeling for intrusion detection. Technical Report CUCS-002-00, Computer Science, Columbia University, 2000.

[4] Mukkamala S and Sung A H. Identifying significant features for network forensic analysis using artificial intelligent techniques. *International Journal of Digital Evidence*, 2003, 1(4): 1-17.

[5] Srilatha Chebrolu, Ajith Abraham, and Johnson P. Thomas. Feature deduction and ensemble design of intrusion detection system. *Computer & Security*, 2005, 24(4): 295-307.

[6] 邹涛，孙宏伟，田新广，李学春. 入侵检测系统中两种审计数据缩减技术的比较与分析.计算机应用，2003, 23(7): 13-17.
Zou Tao, Sun Hong-wei, Tian Xin-guang, and Li Xue-chun. Comparison and analysis of two audit data reduction methods for intrusion setection system. *Computer Applications*, 2006, 23(7): 13-17.

[7] Lippmann R, Haines J W, Fried D J, Korba J, and Das K. The 1999 DARPA off-line intrusion detection evaluation. *Computer Networks* , 2000, 34(4): 579-595.

田俊峰：　男，1965 年生，教授，研究领域为信息安全、网络技术.
王惠然：　女，1982 年生，硕士生，研究方向为信息安全.
傅　玥：　女，1977 年生，硕士生，研究方向为信息安全.