文章编号：2095-2163(2019)01-0178-06 中图分类号：TP393.08 文献标志码：A

即时通讯流量检测与分析

胡 阳，余翔湛，李 凯
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：本文从数据包数量受限条件下的流量检测和基于 Petri 网的协议分析两个方面进行即时通讯流量的研究。通过互信息分析对比不同即时通讯流量数据包数为流量分类提供的信息量差别，利用统计学检验方法以及基于混淆矩阵的分类性能评价方法对机器学习分类器在不同包数下对即时通讯流量的分类的情况进行分析，得到各种机器学习分类器达到最佳分类状态时用于检测的包数量。基于 Petri 网对协议形式化的描述，将协议受到的攻击行为转化为网中插入的新元素，利用矩阵的运算完成协议攻击成功的可能性分析，使得协议安全性的分析有了形式化的方法，避免了人工分析的不确定性和局限性。本文设计并实现了一个即时通讯流量检测分析系统，通过数据包数选取、机器学习分类以及 Petri 网分析，实现包数受限下的即时通讯软件协议数据流识别分类及通讯的消息分析还原。

关键词：流量检测；机器学习；Petri 网；协议分析

The detection and analysis of instant messenger traffic flow

HU Yang, YU Xiangzhan, LI Kai
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 This paper studies the instant messenger traffic flow from two aspects: traffic detection under the condition of limited number of packets and protocol analysis based on Petri nets. Through mutual information analysis, the paper compares the amount of information provided by different packet numbers for traffic classification. Statistical test and performance evaluation based on confusion matrix are used to analyze the result of the classification for getting the best packet number for machine learning classifier. By using formal description of protocol based on Petri network, the paper transforms the attack behavior of the protocol into new elements inserted in the network to analyze the possibility of successful protocol attack by matrix operation. Therefore, this paper designs and implements an instant messenger traffic detection and analysis system, using selection of packet number, machine learning and Petri analysis for instant messenger flow classification and message restore.

【Key words】 traffic detection; machine learning; Petri nets; protocol analysis

0 引 言

随着城市无线局域网和第四代移动通信蜂窝数据网络覆盖度的不断扩大，移动数据网络终端设备用户数目的不断增加[1]，移动数据网络终端设备用户对基于互联网的即时通讯服务的需求已经尤显迫切，即时通讯软件的用户数目正呈现着快速增长的态势，研发即时通讯软件的厂商也在陆续增加，即时通讯软件的数目在近几年间正日渐繁多。不同即时通讯软件由于其功能和性能上的不同设定，目前已有的 TCP/IP 上的应用层协议并不能完全满足开发的需要，因此，研发即时通讯软件的厂商大多选择自行研发私有协议。由于即时通讯软件的种类各异，而各家研发厂商又趋向于使用自行设计的私有协议，导致目前被厂商使用的即时通讯软件私有协议的数目已然迹近庞大，并且各种协议也未能建立统一的标准。

由于网络管理、网络服务质量保证、隐私保护、舆情监控管理、国防等目的推动[2]，需要对即时通讯软件进行有效的监管。考虑到即时通讯软件没有采用公开的已有协议，而是使用了私有协议，并且没有公开协议的具体信息，因此为了对即时通讯软件形成系统完善的监管，就需要对私有协议进行检测识别与分析。

在网络流量较大或能够提供用于流量分类的包数量有限的情况以及在流量的早期识别分类时，不能将每个流的所有数据包都用于流量分类识别，只能使用有限的数据包进行流量识别。在数据包有限的情况下，并不是包的数量越多分类效果越好，因而还要对包的数量与分类效果之间的关系展开研究。

协议的安全性决定了即时通讯服务的安全性，

作者简介：胡 阳(1992-)，男，硕士研究生，主要研究方向：网络安全、流量监测；余翔湛(1973-)，男，博士，教授，博士生导师，主要研究方向：数据安全存储、网络安全、物联网安全等。

收稿日期：2017-06-16

为了挖掘即时通讯服务可能存在的安全漏洞，需要分析协议在攻击下的工作状态。采用形式化的方法对即时通讯协议进行描述，并将攻击行为转化为网系统结构的调整，通过对网的分析来研究讨论协议安全性，可以对协议状态实现充分的检验，从而避免了人工分析的不确定性和局限性。

1　包数受限的流量检测

1.1　数据包互信息分析

在网络流量超出常规的情况下，为每个流提供大量数据包用于分析会导致内存的可观消耗引起系统总体性能的下降。这种情况下，需要对包数限定范围内的数据包进行流量检测分析。在流量传输的早期阶段分类时，为了能够尽早识别现有流量，提高系统的分析识别速度，需要仅通过早期的数据包来研究流量分类[3]。上述 2 个条件下的分类需求都对分类时使用的包数目制定了限制，在包数受限条件下进行流量分类时，需要分析当分类效率达到最高的包数目。

互信息是信息论中提出的用于度量随机变量之间相关性的信息量，即一个随机变量 X 能够给另一随机变量 Y 提供的信息量[4]。互信息已广泛用于特征选取、图像识别、语音识别、生物特征识别等领域。互信息的计算公式为：

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) \quad (1)$$

若变量 X 和变量 Y 并非离散取值，即变量 X 和变量 Y 是连续的，则可将求和转化为二重积分运算：

$$I(X;Y)=\int_{y}\int_{x}p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)\mathrm{d}x\mathrm{d}y \quad (2)$$

本文使用捕获的微信在发送文字消息、发送图片消息、语音呼叫、视频呼叫四种功能下的网络数据流量作为流量检测的样本，将这 4 种功能下的数据流分别作为 4 种不同类型的数据集。在捕获的流量中，只选择载荷不为零的数据包。

互信息分析结果如图 1 所示。文字消息和图片消息的前两个包的互信息比语音呼叫和视频呼叫的包的互信息要高，语音呼叫和视频呼叫的互信息相差不足 0.1。这意味着前两个包并不能为流量识别分类带来足够的信息，而对于 2~4 个数据包的互信息分析中，文字消息和图片消息的结果有了显著的提高。文字消息的 8~9 个数据包为流量识别提供了较高价值的信息，图片消息的 7~8 个数据包为流量识别提供了较高价值的信息，语音呼叫的 6~7 个数据包为流量识别提供了较高价值的信息，视频呼叫的数据包相比其余 3 种功能的数据包较为不同，共有 19~20 个数据包为流量识别提供了较高价值的信息。

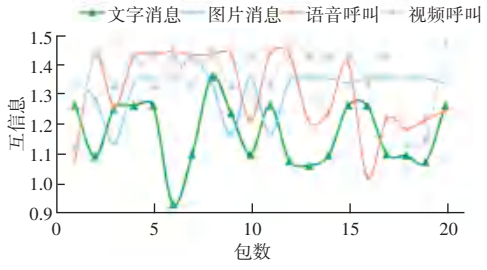

图 1　互信息分析结果

Fig. 1　The result of mutual information analysis

1.2　机器学习分类评价

本文对即时通讯流量进行分类所使用的机器学习分类器有 5 类，分别为贝叶斯分类器、元分类器、规则分类器、决策树以及支持向量机。其中，贝叶斯分类器是基于贝叶斯理论的分类器，目前正流行应用于机器学习领域中。本文使用了贝叶斯网络以及朴素贝叶斯分类器来对流量进行识别。元分类器应用 Bagging 算法[5]以及 AdaBoost 算法[6]来构建分类能力较强的分类器。Bagging 算法首先利用训练集来训练弱分类器，通过多次学习得到分类能力较强的分类器。AdaBoost 算法则通过将多个不同的弱分类能力分类器使用相同的训练集训练之后，集合而形成分类能力强的分类器。规则分类器使用一定策略建立分类规则，之后依据规则对流量进行分类。本文使用了 OneR 和 PART 规则分类器对流量来设计生成识别分类。决策树也可称作统计分类器。C4.5、朴素贝叶斯树、随机森林[7]算法是业界公认的典型决策树算法，本文使用上述决策树算法对流量进行识别分类。支持向量机也可称为有监督的机器学习算法，在分类领域已获得了大规模运用。支持向量机在分类和回归这两类研究中均已获得比较好的效果。本文采用上述 5 类、共 10 种分类器进行流量的分类，分别为贝叶斯网络、朴素贝叶斯、支持向量机、Adaboost、Bagging、OneR、PART、NB 树、C4.5、随机森林。

图 2 为文字消息流量分类的精确度，而与其对应的 ROC 结果即如图 3 所示。由图 2、图 3 中可以看出，前 2~3 个数据包很明显不能为流量分类提供

足够的作用,而随着包的数量从 4 增加到 20,各个分类器的 ROC 结果都呈现出持续的提高。总而言之,大部分的分类器都能达到比较好的 ROC 结果,但是支持向量机和 OneR 分类器的结果却未能臻至理想。在前述的精确性分析中,朴素贝叶斯分类器、霍夫丁分类器、随机森林分类器的分类精确度都不高,但是 ROC 分析的结果却较好,这就说明文字消息数据集中存在不均衡的数据。

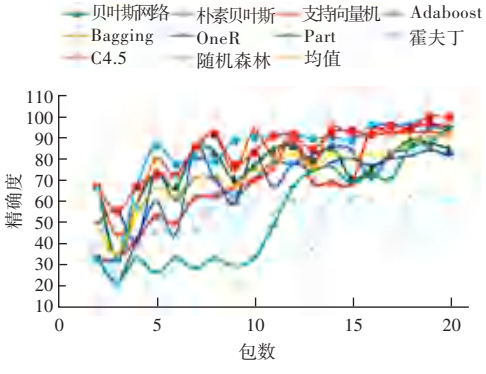

图 2 分类精确度

Fig. 2 The accuracy of classification

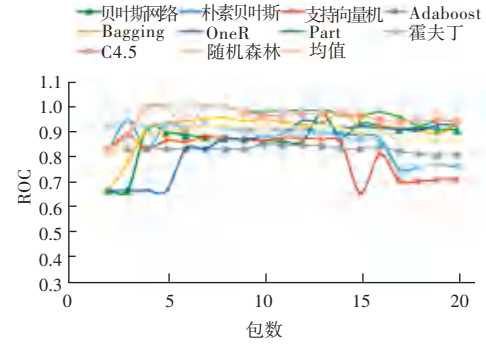

图 3 ROC 分析结果

Fig. 3 The ROC of classification

研究中得到文字消息流量分类精确度检验的弗里德曼检验结果,详见表 1。弗里德曼检验结果显示,包数为 19 的时候,精确度检验结果是最佳的。当对 p 值采取对比研究,并调整 p 值时,包数 10~12 的 p 值比调节后的 p 值要低,包数 15~17 时 p 值也比调节后的 p 值要低,因此包数 10~12 和包数 15~17 是性能最佳的包数值。

为了更好地理解弗里德曼检验结果,本文使用了威尔科克森符号秩检验,可得检验结果见表 2。从表 2 中可以看出包数为 20 的 p 值比包数为 19 的 p 值高不超过 0.05,因此可以得出如下结论,即:包数为 20 时候的分类效果不会比 19 有显著的差别。

表 1 弗里德曼检验结果

Tab. 1 Freedman test result

包数	排名	p 值	调节后 p 值
2	14.967 7	2.518 7	4.029 8
3	17.722 7	1.782 0	3.207 6
4	15.727 7	4.369 4	7.427 9
5	13.511 1	7.155 0	1.073 3
6	11.083 3	7.922 1	1.109 1
7	10.977 7	8.624 3	0.001 0
8	10.977 7	9.362 6	0.001 0
9	13.299 9	1.310 7	1.703 9
10	8.550 0	0.003 8	0.030 1
11	8.550 0	0.001 7	0.015 5
12	6.792 4	0.030 3	0.151 5
13	10.133 3	4.152 0	0.004 2
14	5.594 4	0.102 2	0.306 5
15	8.359 9	0.004 8	0.033 8
16	7.314 9	0.017 2	0.103 0
17	6.270 0	0.051 3	0.205 0
18	3.229 9	0.486 5	0.972 9
19	1.667 7	—	—
20	3.166 6	0.511 1	0.972 9

表 2 威尔科克森符号秩检验结果

Tab. 2 Wilcoxon signed rank test results

与包数 19 对比	R^+	R^-	p 值
2	1.000	27.00	0.027
3	22.00	14.00	0.570
4	47.00	7.50	0.041
5	36.00	0.00	0.011
6	50.50	4.50	0.018
7	44.00	1.00	0.011
8	47.50	7.50	0.041
9	44.00	1.00	0.011
10	54.00	1.00	0.007
11	54.00	1.00	0.007
12	50.00	5.00	0.022
13	54.00	1.00	0.007
14	54.00	1.00	0.007
15	54.00	2.50	0.007
16	52.00	1.00	0.110
17	54.00	1.00	0.007
18	54.00	1.00	0.007
20	54.00	1.00	0.007

从上述分析结果可以看出，即时通讯软件早期的通讯数据包携带了足够的信息，可供流量识别分类，除了支持向量机和 OneR 分类器以外，其余机器学习分类器都能通过早期的流量数据包进行有效的流量识别分类。但是前 3 个数据包在流量早期分类识别时不足以提供足够的信息，只使用 3 个数据包进行分析无法得到理想的分类结果。通过分类精确度的分析结果，可以看出由于数据集中的不平衡数据，有些情况下各种分类器能够达到很高的分类能力，而有些情况下分类能力就并不显著。支持向量机和 OneR 分类器在非零数据包增加的情况下，分类能力也未见到可观改善，这 2 种分类器的性能表现与其余分类器的表现尤为不同，对于即时通讯软件的早期流量分类识别来说并非有效选择。在被测试的 10 种分类器中，随机森林分类器的表现最为突出，是比较适合进行即时通讯软件流量分类识别的分类器。

2　基于 Petri 网的协议分析

Petri 网[8]可用三元组来表示，网 $N=(P,T;F)$ 中，P 被称为库所集，P 中的元素为库所，T 被称为变迁集，T 中的元素为变迁，F 被称为流。对于网 $N=(P,T;F)$，正整数集 N，ω 为无穷，且 $\omega=\omega+1=\omega-1=\omega+\omega$，$K:P\rightarrow N\cup\{\omega\}$ 被称为网的容量函数。对于网 $N=(P,T;F)$，非负整数集 N_0，$M:P\rightarrow N_0$ 称为 N 的一个标识的条件是 $\forall p\in P:M(p)\leqslant K(p)$。将 (N,M) 称为标识网。对于网 $N=(P,T;F)$，$W:F\rightarrow N$ 称为网 N 上的权函数，对 $(x,y)\in F$，$W(x,y)=W((x,y))$ 被称为 (x,y) 上的权。

将协议对应的 Petri 网和攻击过程用矩阵进行表示，Petri 网的运行过程就可以转化为矩阵的乘法运算。协议在攻击前的状态 S_0 对应的 Petri 网标识为 M_0，协议攻击成功的状态对应的 Petri 网标识为 M_c，根据 Petri 网标识可达的充分必要条件，如果存在序列 $M_0[Y_0]>M_1[Y_1]>M_2[Y_2]>M_3\cdots M_n[Y_n]>M_c$，其中 n 为正整数，即这一序列是个有限的序列，针对协议的攻击是可以成功的。

对于网 $\Sigma=(P,T;F,M)$，其结构主要为各个库所与变迁之前是否存在流关系，以及流关系的权值，则网中的矩阵数值表示了库所与变迁之前的边的存在性及其权值。设 $n=|T|$，$m=|P|$，则矩阵 $A=[a_{ij}]_{n\times m}$ 可表示 Petri 网的库所与变迁的关系。矩阵中第 i 行 j 列的元素 $a_{ij}=a_{ij}^{+}-a_{ij}^{-}$，$a_{ij}^{+}=W(t_i,p_j)$，$a_{ij}^{-}=W(p_j,t_i)$，则 $a_{ij}=W(t_i,p_j)-W(p_j,t_i)$。

对于网的攻击前标识 M_0 与攻击后标识 M_c，若存在一个矩阵 $\boldsymbol{X}=[x_{k1}]_{n\times 1}$，满足 $M_c=M_0+\boldsymbol{A}^{\mathrm{T}}\times\boldsymbol{X}$，则表示 M_c 可以从 M_0 到达。其中 $\boldsymbol{A}^{\mathrm{T}}$ 表示矩阵 $\boldsymbol{A}$ 的转置，即 $\boldsymbol{A}^{\mathrm{T}}=[a_{ji}^{'}]_{m\times n}$，其中 $a_{ji}^{'}=a_{ij}$。$\times$ 表示矩阵乘法，$\boldsymbol{A}^{\mathrm{T}}\times\boldsymbol{X}$ 的结果是一个矩阵，行数为 m，列数为 1，设 $\boldsymbol{Y}=\boldsymbol{A}^{\mathrm{T}}\times\boldsymbol{X}$，则 $\boldsymbol{Y}=[y_{l1}]_{m\times 1}$，$y_{l1}=\sum_{h=0}^{n}a_{lh}^{'}x_{h1}$。对于协议的攻击被转化为了对 Petri 网标识可达情况的判断，网标识可达情况的判断又可以转化为对等式能否成立的判断。因此不需要求出矩阵 $\boldsymbol{X}$ 的具体值，只需要判断矩阵 $\boldsymbol{X}$ 的存在性，就可以得知对于协议攻击是否能够成功。

对于等式 $M_c=M_0+\boldsymbol{A}^{\mathrm{T}}\times\boldsymbol{X}$，可以将其转化为如下方程组：

$$\begin{cases}M_{c11}-M_{011}=a_{11}^{'}x_{11}+a_{12}^{'}x_{21}+a_{13}^{'}x_{31}+\cdots+a_{1n}^{'}x_{n1}\\ M_{c21}-M_{021}=a_{21}^{'}x_{11}+a_{22}^{'}x_{21}+a_{23}^{'}x_{31}+\cdots+a_{2n}^{'}x_{n1}\\ M_{c31}-M_{031}=a_{3n}^{'}x_{11}+a_{32}^{'}x_{21}+a_{33}^{'}x_{31}+\cdots+a_{3n}^{'}x_{n1}\\ \cdots\\ M_{cm1}-M_{0m1}=a_{m1}^{'}x_{11}+a_{m2}^{'}x_{21}+a_{m1}^{'}x_{31}+\cdots+a_{mn}^{'}x_{n1}\end{cases}\tag{3}$$

设 $\boldsymbol{\alpha}_i=[a_{ki}^{'}]_{m\times 1}$，$\boldsymbol{\beta}=[M_{ck1}-M_{0k1}]_{m\times 1}$，其中 $k\in\{1,2,3,\cdots,m\}$，则上述方程组可以表示为：

$$\boldsymbol{\beta}=x_{11}\boldsymbol{\alpha}_1+x_{21}\boldsymbol{\alpha}_2+x_{31}\boldsymbol{\alpha}_3+\cdots+x_{n1}\boldsymbol{\alpha}_n\tag{4}$$

若存在不全为零的 k_i，$i\in\{1,2,3,\cdots,n\}$，使得：

$$\boldsymbol{\beta}=\sum_{i=1}^{n}k_i\boldsymbol{\alpha}_i=k_1\boldsymbol{\alpha}_1+k_2\boldsymbol{\alpha}_2+k_3\boldsymbol{\alpha}_3+\cdots+k_n\boldsymbol{\alpha}_n\tag{5}$$

则方程组的解是存在的。

上述的 Petri 网矩阵运算分析的过程，将协议转化为网的形式，攻击行为也转化成了网的形式，将协议承受攻击的情况转化为网运行情况的分析。采用形式化方式分析协议的攻击情况，一方面可以使得分析过程更加清晰，分析结果更加明确，另一方面也可以避免非形式化分析协议时状态检验不全导致的分析不全面的可能性。矩阵运算是一种便于计算机操作的运算，同时可以使用 GPU 进行加速处理，在分析的性能上具有较好的表现[9]。

3　流量检测分析系统

流量检测分析总体结构如图 4 所示。系统共分为 6 个模块，总控制模块负责系统的整体运行控制。配置管理模块将读入并配置各模块运行参数或将更改后的配置信息存入配置文件中供下次使用。包处

理模块向内为系统提供待处理的数据包,向外负责与外部系统进行待处理数据包的获取或响应数据包的传递工作,包处理模块为系统中需要使用通讯流量的模块指定了统一的数据包接口。受限包数流量分类模块使用前文所述的即时通讯流量分类方法,利用有限数目的数据包进行流量的分类识别。Petri 网分析模块对协议的攻击行为进行模拟计算,通过网矩阵运算解存在性的分析进行网攻击成功可能性的判断。内容还原模块为将即时通讯流量中传输的语音、图片等实体信息还原出来。

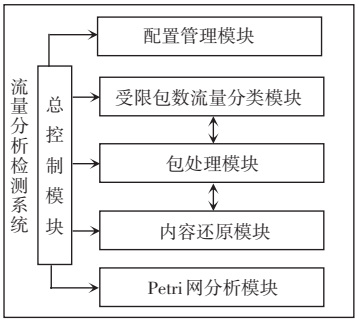

图 4 流量检测分析系统

Fig. 4 Traffic detection and analysis system

系统测试采用捕获的 CocoVoice 即时通讯软件的通讯数据包进行处理,流量检测分析系统读入捕获的包含有 CocoVoice 文字、语音、视频呼叫、音频呼叫等 4 种功能的流量数据,对流量进行识别分类,并从分类后的流量中提取出实际传输的消息实体。为了测试系统对包数选取的效果,限制系统用于分类可以使用的包数,与其相对比的是直接使用分类器进行分类,包数限制为从 2~20,共进行 19 次测试,对每次分类后的分类结果开展精确度计算。图 5 即为系统对流量进行分类的精确度结果。结果显示,本系统能够根据可用于分类的最大包数,动态地调整用于分类的包数目,保证系统整体分类精确度维持在一个比较稳定的水平。而直接使用分类器进行分类,分类器的分类能力受到包数目变化的影响较大,无法达到稳定的分类效果。

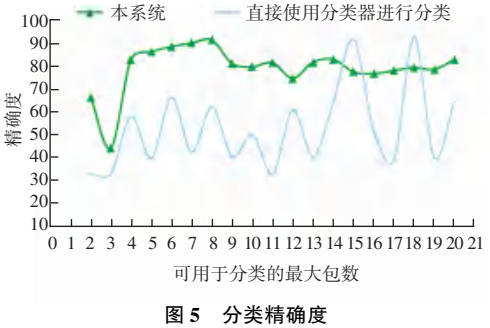

图 5 分类精确度

Fig. 5 The accuracy of classification

图 6 为系统提取出的消息实体,说明系统成功地识别了 CocoVoice 即时通讯软件的流量,内容还原模块将 CocoVoice 即时通讯流量中传输的实际消息图片还原出来,并保存成为图片文件。

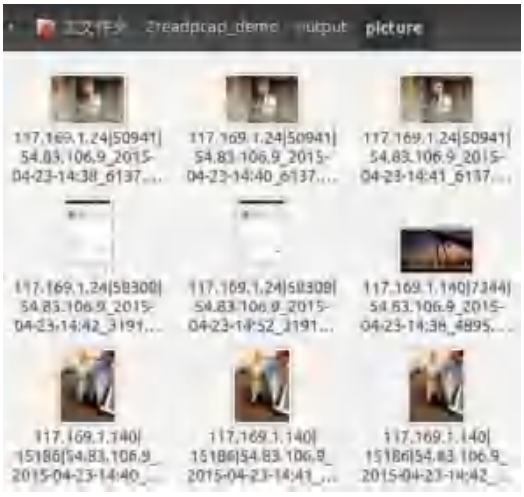

图 6 系统提取出的消息内容

Fig. 6 Message content extracted by system

4 结束语

本文针对即时通讯软件流量,从识别分类与协议分析两个方面进行研究。在流量识别分析方面,提出了包数受限条件下的流量分类识别方法,通过互信息分析计算不同数量的数据包,为流量分类提供信息量,从信息量的角度来对包数受限条件下的流量分类提供包数选择的依据。利用弗里德曼检验和威尔科克森符号秩检验两种统计学检验方法以及基于混淆矩阵的分类性能评价方法对机器学习分类器的分类能力进行分析,并得出使各种机器学习分类器达到最佳分类状态时用于检测的包数量,从各种分类器与在不同包数下的分类性能角度,得出了有限包数下的流量分类的分类器以及包数选取。在协议分析方面,本文提出了基于 Petri 网的即时通讯协议攻击可能性分析方法,使用 Petri 网描述即时通讯协议系统,将攻击者的攻击转化为插入 Petri 网中的元素,通过分析网标识的可达性对攻击的可能性进行判断。本文提出的包数受限条件下的流量分类识别方法在有限的数据包情况下,能够选择效率最高的数据包数目,使得分类器的分类能力与供分类的数据包数达到最佳的匹配。基于 Petri 网的协议分析使用了便于计算机计算的矩阵运算,通过形式化的分析对攻击的可能性进行判断,避免了人工分析的不全面性。

(下转第 187 页)