文章编号:2095-6134(2018)04-0521-08

一种面向未知链路帧的格式特征提取与分类算法*

薛开平†,柳 彬,李 威,洪佩琳

(中国科学技术大学电子工程与信息科学系,合肥 230026)

(2017年4月17日收稿;2017年5月27日收修改稿)

Xue K P, Liu B, Li W, et al. A format feature extracting and classifying algorithm for unknown data link frame[J]. Journal of University of Chinese Academy of Sciences, 2018,35(4):521-528.

摘 要 随着通信网络的发展,私有协议被广泛应用。缺乏必要先验知识时,现有面向已知协议的解析工具无法获取私有协议数据承载的信息。获取私有协议数据承载的信息的前提是正确实现协议格式特征提取与数据分类。基于协议格式一般规律,提出一种针对私有链路协议的未知帧格式特征逆向提取与分类算法。通过链路帧预编码、固定域挖掘从帧样本集合提取帧格式特征并计算特征向量,最后基于特征向量加权欧氏距离对链路帧分类。测试结果表明,该算法能够有效提取帧格式特征,正确实现链路帧的提取和分类。

关键词 私有协议;未知链路帧;格式特征;分类

中图分类号:TP393 **文献标志码**:A **doi**:10.7523/j.issn.2095-6134.2018.04.015

A format feature extracting and classifying algorithm for unknown data link frame

XUE Kaiping, LIU Bin, LI Wei, HONG Peilin

(*Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei* 230026, *China*)

Abstract With the rapid development of communication network, private protocol is widely adopted. Without necessary prior knowledge, the existing analyzing tools for the open protocols cannot be used for obtaining the information from the private protocol data. To get the information from the private protocol data, one has to extract the protocol format feature and classify the protocol data correctly. Based on the general rules of protocol format, a format feature reverse extracting and data classifying algorithm was proposed for unknown data link frame. By data link frame precoding and fixed-field mining, the frame format features can be extracted from the frame sample set and the feature vectors can be calculated. Finally, the data link frames are classified based on the weighted Euclidean distances between the feature vectors. The test results show that the proposed method can be used to extract the protocol format features effectively and to correctly classify the data link frames by using format features.

Keywords private protocol; unknown data link frame; format feature; classification

*国家自然科学基金(61379129)和中国科学院青年创新促进会人才基金(2016394)资助

†通信作者,E-mail:kpxue@ustc.edu.cn

经过多年的研究和发展,面向已知协议的识别分析技术不断进步,并且有现成的协议识别解析工具,如 Wireshark、Tcpdump 等,能有效完成已知协议的识别和分析工作。然而随着通信网络的发展,为保障通信的安全性,在军事、商业以及部分民用的场景中,越来越多的网络通信基于非公开的私有协议进行。面对私有协议数据,由于缺乏必要的先验信息,现有的协议分析工具无法识别私有协议数据所属的协议类型,更无从进一步获取数据承载的有用信息。针对这种情况,研究人员越来越多地采用协议规范逆向挖掘[1]的手段实现私有协议信息的获取。面向未知链路帧的格式解析与分类旨在寻找一种可行的协议格式逆向分析方法,从缺乏先验信息的未知二进制链路帧数据中提取链路帧的格式特征,基于格式特征对二进制链路帧进行分类,为后续协议信息的进一步详细解析提供支撑。

针对协议格式逆向解析问题,2004 年启动的协议信息工程(protocol informatics project)[2]将生物信息学中用于基因搜寻的多序列比对算法引入到报文格式信息挖掘中,通过序列比对识别报文中的固定字段与可变字段,在此基础上获得目标报文的结构信息,其不足在于需要大量报文样本,且对于结构复杂的报文其格式挖掘的效率与准确率较低。考虑到特定条件下能获取的报文样本数量有限,在降低对样本数量要求的前提下,文献[3]提出基于一定量先验信息识别报文结构中的特殊字段,如用户参数、状态标志、长度字段等动态字段的协议逆向分析方法,其局限性主要在于字段识别效果受限于先验知识的丰富程度。文献[4]提出一种以递归聚类为基础的格式与语义挖掘方案,采用基于类型的多序列比对算法对报文进行聚类,再根据识别出的各个字段进行语义推断并基于此进行后续报文内容的进一步挖掘。文献[5]利用协议报文中不同字段取值变化范围的区别,提出一种基于协议报文字段的内容变化分布特征的协议格式挖掘方案。文献[6]基于序列模式挖掘提取协议关键词序列,发现具有时序关系的关键词序列集合,识别协议的固定域和可变域,并结合密文随机性特征解析安全协议格式。文献[7]提出一种语义层次的协议格式提取方法,将基于中间语言的动态污点分析思想应用于协议逆向分析领域,依据字段语义的不可分割性,在语义层次实现协议字段识别。文献[8]针对现有协议格式逆向方法在现实中复杂语义环境下存在的逆向准确度不高的问题,利用“协议结构与代码数据结构之间的协同映射”这一规律,提出识别数据结构的协议格式逆向方法。文献[9]将隐式马尔科夫模型用于应用层网络协议报文建模,刻画报文的字段跳转规律和字段内的马尔科夫性质,基于最大似然概率准则确定协议关键词的长度,自动重构协议的报文格式。文献[10]通过报文分类、多序列对比和特定域识别等多个阶段,实现网络协议报文结构的自动识别。

在未知协议数据分类方面:文献[11]提出一种基于多模式匹配思想的网络视频流发现与分类算法;文献[12]基于载荷部分字节的编码,实现对混合流量按应用类型进行有效分类;文献[13]利用载荷大小、分组的信息熵等网络流量特征,利用最短划分距离的方法构建分类模型,实现在线网络流量分类;文献[14]基于隐式马尔科夫模型,分别结合报文长度和到达时间间隔两种行为模式对应用协议进行分类;文献[15]在文献[14]的基础上利用 k 均值聚类方法建立一种混合模型,能准确地识别同时具有多种行为模式的应用协议,分类准确率较文献[14]提升超过 30% 。然而文献[11-15]都只适用于对网络层以上的协议进行识别分类,无法处理未知的链路层比特流数据。

本文基于协议格式特征的一般规律,通过链路帧预编码、固定域挖掘从帧样本集合中提取协议格式特征并计算格式特征向量,在此基础上计算链路帧与格式特征的特征向量之间的加权欧氏距离,基于欧氏距离实现对未知链路帧的有效分类。本文的创新点包括:1)提供一种通过预编码与固定域挖掘提取帧格式特征的方法,保证帧格式特征提取的完整性;2)提出一种基于特征向量加权欧氏距离的链路帧分类方式,确保链路帧分类的准确性。

1 协议格式特征

通常情况下,对一种链路帧而言,当其承载的上层负载类型确定之后,其格式特征可抽象概括成如图 1 所示的固定域与可变域间隔出现的形式。

其中,固定域是指在链路帧中出现位置固定,且长度和比特内容均固定的比特字段。这些固定域一般以帧头同步序列、协议版本号、长度字段和

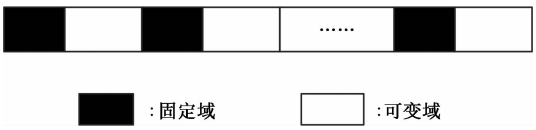

图 1　协议格式特征

Fig. 1　Protocol format feature

端口号等形式出现。以以太网中传输 http 数据的以太帧为例(假设网络层为 IPv4 协议,TCP 的源端口号为 http 默认端口 80 端口),其固定域与可变域间隔的格式特征如表 1 所示。

表 1 中 7 * 0xAA + 0xAB 为以太帧头部的同步序列,用于标识一个以太帧的开始,在所有以太帧中都相同。固定域 0x0800 为以太帧头部标识负载类型为 IPv4 数据包的类型码,0x45 为 IPv4 头部的版本号和长度字段。1 byte 长的固定域 0x06 为 IPv4 头部标识传输层协议类型为 TCP 协议的协议字段,最后一个固定域为标识 http 协议的 TCP 端口号 80(0x0050,表 1 中假设为源端口号)。通常情况下,在以太网中传输 http 数据的以太帧均具有表 1 所示的固定域与变化域相间的格式特征。

表 1　传输 http 数据的以太帧格式特征

Table 1　Features of the Ethernet frame format that transmits the http data

长度/byte	8	12	2	1	8	1	10	2	可变
是否固定	是	否	是	是	否	是	否	是	否
内容	7 * 0xAA + 0xAB	地址域	0x0800	0x45	IP 头	0x06	TCP 头	0x0050	TCP 数据

基于以上分析我们认为,如果帧的链路协议类型确定,同时传输的上层负载确定,那么其对应的固定域和变化域相间的格式特征是唯一确定的。未知链路帧格式特征提取与分类的目标是在缺乏先验知识的情况下,提取链路帧的格式特征,按格式特征对链路帧进行分类。格式特征提取的核心是帧格式中固定域的挖掘,即帧内位置固定、比特内容固定的序列的挖掘。

2　协议格式特征提取

提取协议格式特征时,为降低特征提取的时间开销,首先对待分析的协议数据集按一定的比例进行抽样,从抽取的数据样本集合中提取协议格式特征。

为提取协议格式特征,必须挖掘格式特征中的固定域。固定域在链路帧中出现位置固定,且长度和比特内容均固定,因此挖掘固定域必须同时考虑比特序列在链路帧中的"位置"和"比特内容"两个因素,排除其他位置随机出现的内容相同的比特序列的干扰。

2.1　链路帧预编码

为便于在固定域挖掘时同时考虑序列的"位置"和"比特内容",在挖掘之前需要对链路帧原始比特序列进行预编码处理。预编码以字节为基本单位进行,每个字节经编码映射为 5 个字符。前 3 个字符为该字节偏移的 16 进制表示,从零开始计数,其中字节偏移是指该字节在所属链路帧的位置序号,第 4、5 个字符为该字节的 16 进制表示值[6]。表 2 给出对一条链路帧的原始比特序列按此方式进行预编码的示例。

表 2　链路帧预编码示例

Table 2　Example of link frame precoding

编码前	00000000	00010001	10001111	.....
编码后	00000	00111	0028F	.....

对链路帧样本编码完成之后,如果统计发现一个 5 字符元素(即对一字节编码所得值)在编码后的样本集合中出现 m 次,即表明特定 8 比特序列在原始样本集的链路帧的特定位置出现 m 次。后续的固定域挖掘将一个字节编码后所得的 5 个字符作为一个整体,以 5 字符为基本单位进行。

2.2　固定域挖掘

链路帧样本的预编码完成之后,即可设定阈值参数在样本集合中进行固定域的提取。

由于样本集合中的链路帧负载类型不尽相同,对于负载类型不同的链路帧,其固定域与可变域间隔出现的格式特征也不相同,具体体现在负载类型不同的链路帧,其格式特征中的固定域数量、固定域位置以及固定域比特内容之间的差别。表 3 所示为承载的负载类型分别为 ARP 数据包和 IPv4 数据包(假设应用层协议为 http)的以太帧格式特征(为便于表示,表中仅按序以预编码之后的形式列出两种帧格式中的固定域)。

表3 以太帧格式特征

Table 3 Features of Ethernet frame format

类型	固定域1	固定域2	固定域3	固定域4	固定域5	固定域6	固定域7	固定域8
Ethernet-ARP	同步序列	00C08	00D06	00E00	00F01	无	无	无
Ethernet-IPv4-TCP	同步序列	00C08	00D00	00E45	00F00	01706	02400	02550

假设我们的链路帧样本中只存在表3中两种类型的链路帧,与帧格式中的可变域相比,样本集合中链路帧的固定域会呈现出固定不变(如固定域1的同步序列)或低频变化(如固定域4和5)的特征。固定域的低频变化特性指的是,在帧样本集合中,在链路帧的某一位置(如固定域4的位置00E)处出现的固定域种类较少(表3中仅00E00与00E45两种)且特定固定域出现频率较高的特性(在样本容量n一定时,固定域种类数k较少必然使得其中至少一种固定域出现频率较高)。

基于以上分析,我们在从预编码后的链路帧样本集合中挖掘固定域时设定如下两个阈值参数:

1)支持度阈值sup

设某个元素(即编码后的5字符基本单位)在帧样本集合中出现次数为m,样本容量为n,当m/n不低于sup时认为该元素为可能的固定域(频繁元素)。

2)种类数阈值k

在链路帧样本集合中,某频繁元素所处位置处的元素种类数k_0不超过k时,认为该频繁元素为帧格式中的固定域。

帧格式中的固定域在包含多种类型负载的帧样本集合中除呈现低频变化特性,不同位置的固定域出现的频率也不完全相同,一般帧格式中位置靠后的固定域在样本集合中出现频率会低于位置靠前的固定域。以表4的样本集合为例,假设样本容量为1 000,上层协议类型及数量如表4所示。

表4 链路帧样本集合示例

Table 4 Example of a link frame sample set

类型/数量	固定域1	固定域2	固定域3	固定域4	固定域5	固定域6
Ethernet-IPv4-UDP/300	同步序列	00C08	00D00	00E45	00F00	01717
Ethernet-IPv4-TCP/400	同步序列	00C08	00D00	00E45	00F00	01706
Ethernet-IPv4-ICMP/300	同步序列	00C08	00D00	00E45	00F00	01701

对Ethernet-IPv4-UDP类型来说,位置靠前的固定域1~4为3种类型所共有,在帧集合中出现的频率均为1 000,而固定域01717为Ethernet-IPv4-UDP类型特有,只出现300次。在设定的支持度阈值参数较高时,这种在格式特征中出现位置靠后且频率较低的固定域往往会因为达不到阈值要求而不能被挖掘出来,导致最终提取出的格式特征不完整。

为解决上述问题,提取出尽可能完整的帧格式特征,将第一次挖掘出的固定域缓存在一个固定域集合中,同时记录各固定域所属的帧,将包含同一固定域的帧作为一个样本子集,将原始帧样本集合按包含固定域的不同划分为若干子集,帧集合的一个子集即对应一个固定域。将样本分集之后,按设定的阈值参数(sup,k)在各子集中分别继续进行固定域挖掘,将挖掘出的固定域继续缓存在固定域集合中,同时记录各固定域所属的帧,并依此将各帧集合再次分为若干个子集。重复此过程,直到不能再挖掘出满足阈值(sup,k)的固定域为止。

2.3 协议格式特征的生成

按2.2节所述方式挖掘固定域,所得的是若干固定域组成的无序集合,为了从挖掘出的固定域中获得完整的协议格式特征,在从链路帧样本集合中挖掘出固定域之后,需要按照一定的方式将固定域组织成树结构。

假设存在如图2所示的固定域挖掘过程,在此过程中,第一次从链路帧样本集合(记为set1)提取出固定域000AA、010AB和010AC,其中000AA为所有链路帧样本共有,包含固定域010AB与010AC的帧分别构成子集set2、set3。在set2中继续挖掘固定域得到020EF(对应的帧子集为set4)与020AF(对应的帧子集为set5),在set3中继续挖掘固定域得到03156(对应的帧子集为set6)与031EF(对应的帧子集为set7)。

可根据挖掘的先后顺序与固定域所属帧集合之间的包含关系,将所有固定域组织成如图3所示的树结构。图3树结构中的一条路径即对应一种类型的格式特征。

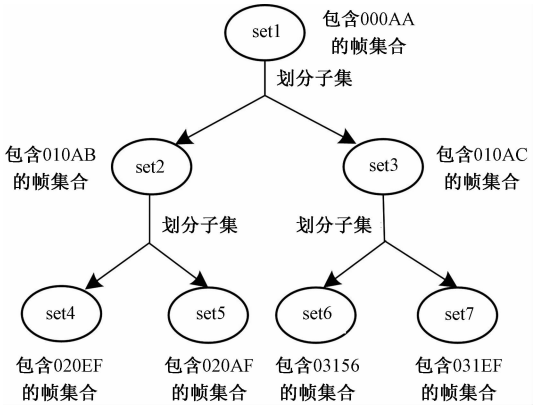

图 2　固定域挖掘与链路帧分集

Fig. 2　Fixed-field mining and link frame separation

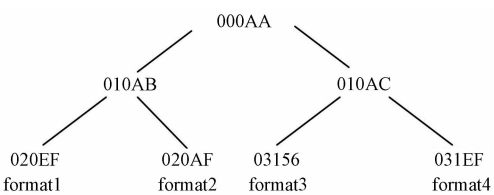

图 3　固定域树结构

Fig. 3　Tree structure of fixed-field

获得固定域的树结构之后,可为每种格式特征(format_i,即树结构中的路径)建立特征向量,format_i 的特征向量定义为 k 维向量:$\boldsymbol{V}_i = (v_{i1}, v_{i2}, v_{i3}, \cdots, v_{ik})$,其中 k 为树结构中的固定域总数,

$$v_{ij} = \begin{cases} 1, & \text{if pattern}_j \in \text{format}_i \\ 0, & \text{if pattern}_j \notin \text{format}_i \end{cases} \quad j = 1,2,\cdots,k. \tag{1}$$

式中:pattern_j 表示树结构中的一个节点(即一个固定域);v_{ij} 与树结构中的固定域一一对应,当 format_i 包含固定域 pattern_j 时 v_{ij} 取 1,否则取 0。按此定义,图 3 中的 4 种 format 对应的特征向量计算结果如表 5 所示。

表 5　特征向量计算结果

Table 5　Calculation results for the feature vectors

	$\boldsymbol{V}_1$	$\boldsymbol{V}_2$	$\boldsymbol{V}_3$	$\boldsymbol{V}_4$
000AA	1	1	1	1
010AB	1	1	0	0
010AC	0	0	1	1
020EF	1	0	0	0
020AF	0	1	0	0
03156	0	0	1	0
031FF	0	0	0	1

至此,完成链路帧样本集合中帧格式特征的提取,每种结构特征均以特征向量形式表示,后续的帧分类即可基于帧格式的特征向量进行。

3　基于格式特征的分类

为了描述一条链路帧在多大程度上符合从帧样本集合中提取出的一种格式特征,与格式特征的特征向量类似,为链路帧 frame_i 定义特征向量:

$$\boldsymbol{Fra}_i = (f_{i1}, f_{i2}, f_{i3}, \cdots, f_{ik}). \tag{2}$$

式中:k 为树结构中的固定域总数;与 v_{ij} 类似,$f_{ij}(j = 1,2,\cdots,k)$ 与树结构中的固定域一一对应,当 frame_i 包含固定域 pattern_j 时 f_{ij} 取 1,否则取 0。

在计算出链路帧 frame_i 的特征向量之后,为将链路帧 frame_i 划分到与之格式最接近的格式类型,定义 frame_i 与格式 format_j 之间的距离为 $\boldsymbol{Fra}_i$ 与 $\boldsymbol{V}_j$ 之间的加权欧氏距离:

$$\text{dis}(\boldsymbol{Fra}_i, \boldsymbol{V}_j) = \sqrt{\sum_{i=1}^{k} \alpha^{n-m} (f_{ik} - v_{jk})^2}. \tag{3}$$

式中:n 为固定域树结构深度;m 为 f_{ik}(或 v_{ik})对应固定域(节点)在树结构中所处的层次;α 为大于 1 的常数。由于提取固定域时采取的是对样本不断分集分步挖掘的方式,因此位置越靠后(对应树结构中层次越深)的固定域提取时所用样本数越少,其可信度逐层降低,α^{n-m} 的作用是使得位置靠前的固定域在计算距离时所占权重增大,位置越靠后所占权重减小以提高分类准确率。

计算出链路帧 frame_i 与各个 format 之间的距离之后,比较距离大小,将链路帧 frame_i 划分到与其距离最小的 format 类。

4　实验

本文采用实际数据对算法进行测试,使用 Qt Creator 3.1.0 平台实现所提出的算法,硬件配置为:Intel i5-2450 M,2.5 GHz,双核 CPU,RAM 4 GB。

4.1　算法有效性验证

为验证算法的有效性,实验采用 Wireshark 实际抓获的 Ethernet 数据集,大小为 35.6 MB,共包含 12 040 个 Ethernet 帧,数据集中的各层协议组成如图 4 所示。

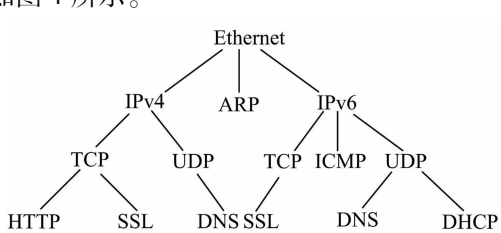

图 4　数据集的组成

Fig. 4　Composition of the data set

4.1.1 格式特征提取

从数据集中抽取1 200个Ethernet帧作为样本,分别设定支持度阈值参数 sup = 0.45,种类数阈值参数 $k=2$ 在帧样本集合中进行固定域提取,所得固定域树结构如图5所示。

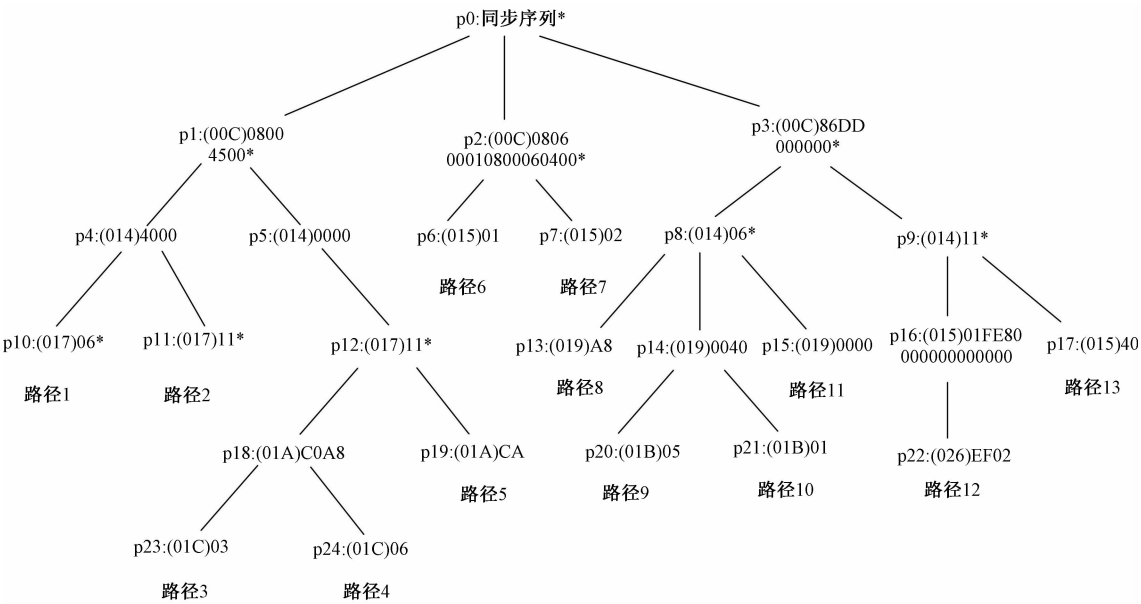

图5 从样本集合中提取的固定域树结构

Fig. 5 Fixed-field tree structure extracted from the sample set

为方便表示,图中将位置连续的固定域合并后表示,如固定域 p1:(00C)08004500,是由位置连续的4个固定域00C08,00D00,00E45以及00F00合并而成,合并后保留第一个固定域的位置信息00C;尾部带*的固定域,如p0(帧头同步序列),p1(IPv4类型码及头部版本号字段),p2(ARP类型码及头部固定字段),表示正确提取出的固定域,这些固定域存在于相应类型链路帧格式特征中,尾部不含*的固定域是由于样本集的选取与参数设置原因导致的"伪固定域",链路帧格式特征中不含这些固定域。自根节点p0到每个叶子节点有且仅有一条路径,图5中将所有路径按在树结构中从左到右的顺序依次记为路径1,…,路径13。

图5中,重点关注将正确提取出来的固定域按序组织后可表征的帧格式类型,如表6所示。从表6可以看出,在提取出的图5所示的树结构中,提取正确的固定域可按序组织成5种有效的帧类型格式特征。

4.1.2 链路帧分类

基于4.1.1提取的固定域树结构,按第3节中的方式计算格式特征与链路帧的特征向量,取 $\alpha=1.2$,根据特征向量间的加权欧氏距离确定链路帧所属的路径,对链路帧进行分类,分类结果如表7所示。

表6 固定域组合与帧类型的对应关系

Table 6 Correspondence between fixed-field combination and frame type

序号	固定域组合	对应帧类型	对应路径
NO.1	p0-p1-p10	Ethernet-IPv4-TCP	1
NO.2	p0-p1-p11	Ethernet-IPv4-UDP	2
NO.3	p0-p1-p12	Ethernet-IPv4-UDP	3,4,5
NO.4	p0-p2	Ethernet-ARP	6,7
NO.5	p0-p3-p8	Ethernet-IPv6-TCP	8,9,10,11
NO.6	p0-p3-p9	Ethernet-IPv6-UDP	12,13

从表7可以看出,基于加权欧氏距离对链路帧进行分类时,虽然树结构中存在部分"伪固定域",但由于其出现位置靠后且分类时取 $\alpha=1.2$,"伪固定域"在分类时所占权重较小,不影响分类的准确性,所有类型的帧均能被正确划分到对应的路径(即帧格式类型)。

4.2 参数设置对算法的影响

4.2.1 阈值参数对算法的影响

在算法设计中,固定域的挖掘涉及阈值参数(sup,k)的设置,为分析参数(sup,k)对算法的影响,实验采用4.1节相同的数据集,调整参数取值观察实验结果。为简化输出结果,未列出设置不同参数时提取出的具体固定域树结构,而是以表格形式给出树结构中的固定域数目、路径数和

表 7　分类结果

Table 7　Classification results

帧类型	格式特征/%				
	Ethernet-IPv4-TCP（路径 1）	Ethernet-IPv4-UDP（路径 2,3,4,5）	Ethernet-ARP（路径 6,7）	Ethernet-IPv6-TCP（路径 8,9,10,11）	Ethernet-IPv6-UDP（路径 12,13）
Ethernet-IPv4-TCP	100.0	0.0	0.0	0.0	0.0
Ethernet-IPv4-UDP	0.0	100.0	0.0	0.0	0.0
Ethernet-ARP	0.0	0.0	100.0	0.0	0.0
Ethernet-IPv6-TCP	0.0	0.0	0.0	100.0	0.0
Ethernet-IPv6-UDP	0.0	0.0	0.0	0.0	100.0

格式特征类型等信息,具体实验结果如表 8 所示(在统计树结构中固定域数目时,挖掘出的一个固定 5 字符元素即作为一个固定域,不对固定域进行合并)。由表 8 可知,随着阈值的降低(支持度参数 sup 降低,种类数参数 k 增大),算法提取出的固定域总数增多,其中包含的正确固定域数目增多,相应的树结构中的路径总数和包含的格式特征类型数也随之增多。通过降低阈值参数可从样本集提取出更丰富的协议格式信息,实现对链路帧更细粒度的分类。

对比 sup = 0.45 时 k 取值分别为 2、3、4 的情形可以发现,降低阈值参数虽然使得正确挖掘出的固定域数增多,但方案提取出的有效特征类型数相同。这是由于因为划分子集导致样本集容量不断减小,使得应用类型特征相关的固定域不能完整挖掘出来,方案未能获得完整的自链路层至应用层的协议格式特征。在这 3 种情形下,方案正确提取出的帧格式特征数以及分类粒度相同。

表 8　阈值参数对算法的影响

Table 8　Influences of threshold parameters on the algorithm

sup	k	固定域总数	正确固定域数	路径总数	有效特征类型数	时间/s
0.45		52	23	13	5	132
0.55	2	44	21	12	4	127
0.65		14	9	5	2	87
	2	52	23	13	5	132
0.45	3	75	26	21	5	163
	4	78	29	22	5	168

降低阈值参数在获取更丰富的协议格式信息的同时,由于固定域数量的增多,使得特征向量长度增大,链路帧特征向量以及基于加权欧氏距离的分类所需计算量增大,使得方案耗时增多。

4.2.2　权重参数 α 对算法的影响

在算法设计中,固定域的挖掘涉及权重参数 α 的设置,为分析参数 α 对算法分类准确率的影响,实验采用 4.1 节相同的数据集,设置 sup = 0.45,k = 2,分别取 α = 1.1、1.0 和 0.9 进行链路帧分类并计算准确率,结果如表 9 所示。

表 9　α 对分类准确率的影响

Table 9　Influences of α on the classification accuracy

α	T/F	Ethernet-IPv4-TCP	Ethernet-IPv4-UDP	Ethernet-ARP	Ethernet-IPv6-TCP	Ethernet-IPv6-UDP
1.1	T	100.0	100.0	100.0	100.0	100.0
	F	0.0	0.0	0.0	0.0	0.0
1.0	T	100.0	100.0	100.0	100.0	100.0
	F	0.0	0.0	0.0	0.0	0.0
0.9	T	58.2	100.0	100.0	91.5	99.9
	F	41.8	0.0	0.0	8.5	0.01

从表 9 可以看出,当 $\alpha \geq 1$ 时,由于固定域树结构中的“伪固定域”在分类时所占权重不高于正确提取的固定域,方案可以实现对链路帧所属类型的正确划分;当 $\alpha < 1$ 时,由于固定域树结构中的部分“伪固定域”在分类时所占权重超过正确提取的固定域,“伪固定域”对链路帧的分类造成干扰,使得部分类型的链路帧的分类准确率降低。因此,方案在对链路帧进行分类时,一般取 $\alpha \geq 1$。

以上实验结果表明,对结构中存在固定域与变化域间隔出现的格式特征的链路帧数据集,方案可正确提取出其中的格式特征并基于格式特征实现对链路帧的有效分类。

5 结束语

在私有协议广泛应用的背景下,对采用私有协议传输的数据,如何在缺乏先验知识的情况下获取数据承载的有用信息是一项重要的研究课题。本文分析协议格式的一般特征,根据协议格式的一般特征设定门限参数从未知链路帧样本集合中提取协议格式特征并计算特征向量,在此基础上利用特征向量之间的加权欧氏距离实现对未知链路帧的有效分类。最后利用实际数据实验验证方案的有效性,实验结果表明方案能实现缺乏先验知识情形下的未知协议格式解析与数据分类,为后续协议详细信息的进一步分析奠定了基础。

参考文献

[1] Narayan J, Shukla K, Clancy T C. A survey of automatic protocol reverse engineering tools [J]. ACM Computing Surveys, 2016, 48(3):1-26.

[2] Beddoe M. The protocol informatics project [EB/OL]. (2004). [2016-09-28]. http://www.4tphi.net/~awalters/PI/PI.html.

[3] Cui W, Paxson V, Weaver N, et al. Protocol-independent adaptive replay of application dialog[C]//Proceedings of the 13th Annual Network and Distributed System Security Symposium. IEEE, 2006: 487-490.

[4] Cui W, Kannan J, Wang H J. Discoverer: automatic protocol reverse engineering from network traces[C]//Proceedings of the 16th USENIX Security Symposium. IEEE, 2007: 199-212.

[5] Trifilo A, Burschka S, Biersack E. Traffic to protocol reverse engineering[C]//Proceedings of the Second IEEE International Conference on Computational Intelligence for Security and Defense Applications. IEEE, 2009: 1-8.

[6] 朱玉娜, 韩继红, 袁霖, 等. SPFPA: 一种面向未知安全协议的格式解析方法[J]. 计算机研究与发展, 2015, 52(10): 2 200-2 211.

[7] 潘皞, 洪征, 周振吉, 等. 语义层次的协议格式提取方法[J]. 通信学报, 2013, 34(10): 162-173.

[8] 任翔宇, 谈诚, 赵磊, 等. 识别数据结构的协议格式逆向推理方法[J]. 武汉大学学报(工学版), 2015, 48(2): 269-273.

[9] 罗建桢, 余顺争, 蔡君. 基于最大似然概率的协议关键词长度确定方法[J]. 通信学报, 2016, 37(6): 119-128.

[10] 李伟明, 张爱芳, 刘建财, 等. 网络协议的自动化模糊测试漏洞挖掘方法[J]. 计算机学报,2011, 34(2): 242-255.

[11] 孙钦东, 郭晓军, 黄新波. 基于多模式匹配的网络视频流识别与分类算法[J]. 电子与信息学报, 2009, 31(3): 759-762.

[12] 王变琴, 余顺争. 未知网络应用流量的自动提取方法[J]. 通信学报, 2014, 35(7): 164-171.

[13] 杨哲, 李领治, 纪其进, 等. 基于最短划分距离的网络流量决策树分类方法[J]. 通信学报, 2012, 33(3): 90-102.

[14] Wright C, Monrose F, Masson G M. HMM profiles for network traffic classification[C]//Proceedings of the 2004 ACM workshop on Visualization and Data Mining for Computer Security, ACM, 2004: 9-15.

[15] Wright C V, Monrose F, Masson G M. Towards better protocol identification using profile HMMs[R]. Technical Report JHU-SPAR051201, Johns Hopkins University, 2005.