基于生物信息的未知二进制协议聚类方法

丛培鑫[1]，李晓慧[2]，王俊峰[1]
(1. 四川大学计算机学院，成都 610065；2. 四川大学网络空间安全学院，成都 610065)

摘　要：协议聚类是协议逆向工程技术中非常重要的一步，针对二进制协议更加透明且满足的协议种类更加广泛的特点，提出了一种基于基因和蛋白质生物信息的二进制协议聚类方法，能够从原始序列角度对大量协议直接进行聚类. 本文方法首先将原始二进制报文转化成四进制基因形式，使用快速聚类方法计算碱基两两组合的 k-seed 值生成距离矩阵，并用 UPGMA 计算最小距离生成树得到初始分簇；其次，将每一簇四进制协议报文转化成十六进制蛋白质链，得到序列更有语义的方式并采用基于改进 mBed 算法的聚类方法将其进行高精度聚类. 通过对已知和未知协议单纯和混合场景下的测试表明，该方法能够对二进制协议实现高效并且高准确率的聚类，具有较高的应用价值.

关键词：未知协议；二进制协议；生物信息学；多序列比对
中图分类号：TP301.6　　**文献标识码**：A　　DOI：10.19907/j.0490-6756.2022.032004

Unknown binary protocol clustering method based on biological information

CONG Pei-Xin[1], *LI Xiao-Hui*[2], *WANG Jun-Feng*[1]
(1. College of Computer Science, Sichuan University, Chengdu 610065, China;
2. School of Cyber Science and Engineering, Sichuan University, Chengdu 610065, China)

Abstract: Protocol clustering is a very important step in protocol reverse engineering technology. Aiming at the characteristics of binary protocols that are more transparent and satisfying a wider range of protocols, a binary protocol clustering method based on genetic and protein biological information is proposed, which can learn from the original sequence Angle to cluster protocols directly. The method firstly converts the original binary message into a quaternary gene form, uses the fast clustering method to calculate the k-seed value of the base pairwise combination to generate a distance matrix, and uses UPGMA to calculate the minimum distance spanning tree to obtain the initial cluster; A cluster of quaternary protocol messages is converted into a hexadecimal protein chain, and the sequence is obtained in a more semantic way. The clustering method based on the improved mBed algorithm is used to cluster them with high precision. Tests under pure and mixed scenarios of known and unknown protocols show that this method can achieve efficient and high-accuracy clustering of binary protocols, and has high application value.

Keywords: Unknown protocol; Binary protocol; Bioinformatics; Multiple sequence alignment

收稿日期：2021-11-30
基金项目：基础加强计划重点项目(2019-JCJQ-ZD-113)
作者简介：丛培鑫(1997－)，山东东营人，硕士研究生，研究方向为网络空间安全. E-mail：297970474@qq.com
通讯作者：王俊峰. E-mail：wangjf@scu.edu.cn

1 引　言

没有网络安全就没有国家安全[1]. 随着网络安全的重要性日益增加,网络中数据的通信量也在与日剧增. 网络协议作为网络通信的核心传递方式,它的种类和结构直接关系到通信的安全性、稳定性和可靠性. 比特流作为网络协议数据传输的一个重要载体,其中传输的二进制协议有着传输效率高、适用场景广泛等特点. 由于未知协议的广泛使用,比特流中 40% 以上的流量数据难以被识别和分析[2]. 因此在没有先验知识的情况下,定性地对未知协议进行分类就成为了协议逆向工程的第一步. 这一步将未知二进制协议聚类到不同集合中,能够降低人工分析对协议规范的依赖[3],同时对进一步推断协议格式、提高网络协议分析的效率和准确率都具有重要意义.

网络协议通常分成两大类,即公有网络协议和私有网络协议. 第一类公有协议包括如绝大部分协议依托的网络层 IP 协议,传输直播流量的 UDP 协议,平时访问网页经常用到的 HTTPS、DNS 协议等. 公有协议往往是已知结构的,体现为字段格式公开,取值约束已知,也是各种标准化组织、网络运营商、网络通信技术方案提供者[4]等制定的公开协议. 第二类协议往往存在于卫星、雷达、无人机等特殊设备,多数企业和组织单位,可穿戴设备使用的蓝牙协议以及各个厂商推出的智能家居系列产品等场景. 特别对于企业来说大量的数据在公网上进行传输,如果不采用私有协议会导致用户和企业的敏感数据易被不法分子拦截分析,损害用户和企业的利益. 在此场景下使用私有协议进行传输,可以在实现数据传输的同时,还能满足私有网络的经济利益、保护商业利益等安全性需求. 而随着各大企业的产品覆盖生活的方方面面,私有协议的数量和种类正在快速增加. 私有协议在传输时因为需要足够安全,往往会按照自己的标准,提出新的协议规范构造协议报文. 这类协议规范由于不会公开,所以对外界来说往往是未知的.

当前公有协议解析工具成熟,但对于新协议的规范分析仍面临较大的困难. 目前现有的协议分析软件仅支持对已知协议的分析,例如 Wireshark 可以解析 2000 余种已知协议,对于协议规范没有公开的各类私有协议则无能为力. 在这种情况下,获取未知协议规范,定性定量地进行协议逆向解析就显得尤为重要.

当前私有协议解析尚不能实现协议分类的全面性. 之前协议规范的挖掘过于依赖人工分析,使得对未知协议的解析工作成为了一项准确率没有保证,时间耗费长且具有挑战性的任务. 例如,开源项目 Samba[5]通过人工分析的方法花了 12 年的时间才基本生成私有协议 SMB[6]的协议规范. 因此,研究自动化的协议逆向分析技术在网络安全领域具有重要意义. 而协议聚类工作能够从原始流量角度出发,将其中相同种类的报文重新组合,成为以数据为基础的协议逆向工程中不可或缺的一步,并逐渐成为协议漏洞挖掘[7]和工控协议安全需求[8]的重要工作. 作为研究对象的网络协议可以分为文本协议和二进制协议两种. 目前的研究通常依赖于文本类协议. Discover[9]模拟了报文解析的层次化过程,首次提出了在协议解析过程中存在一些关键字段决定报文子结构解析方式的假设. 潘璠等[10]提出一种基于递归聚类的协议解析方法,使用报文中的格式标识字段(Format Distinguisher, FD)和文本段提取层次化报文结构. 这些方法和开源工具 NetZob[11,12]一样都依赖于文本类协议,无法解析二进制协议.

相比于文本协议,二进制协议通常以 ASCII 码形式呈现,序列透明且限制条件较少[13],是非常好的协议报文聚类对象,大大降低了协议聚类对协议种类和报文特征的依赖. Yan 等[14]提出基于改进的主成分分析法和改进的密度峰值聚类对二进制协议聚类. 张路煜等[15]提出了应用卷积神经网络的方法进行协议识别. 但是在训练过程中未知协议的种类数量需要给定,缺乏普适性.

生物信息学聚类方法依托于序列数据,简洁紧凑、传输较快[7]、适用场景更广的二进制协议就是很好的聚类目标,并且多序列比对匹配精度高,具备从序列角度解决分类问题的能力. 从 PI[16]项目开始,生物信息学方法中的渐进多序列比对算法就已经用于协议逆向工程领域. 学者改进后又诞生了 Discover、AutoFormat[17]、Tupni[18]等传统方案和 Biprominer[19]、Pro-Decoder[20]、Proword[21]等改进方案. 生物信息学已经逐渐成为解决协议分析问题有效的解决方法之一.

本文提出一种基于生物信息的未知二进制协议聚类方法. 该方法面向协议原始二进制报文序列,不局限于文本协议中的特定字段,可以广泛应用于文本协议和二进制协议. 首先对原始二进制数据包内序列进行预处理,分离出二进制报文序列,

对二进制序列转化成四进制的基因序列进行快速聚类，实现预分簇，最后再次对分好后的每一簇协议序列进行进制转换，将四进制序列转化成十六进制的蛋白质链进行二次高精度聚类，得到协议子集. 本文使用真实环境下的网络流量数据进行了大量实验，实验结果较其他算法在准确率和效率上均有一定的提升.

2　相关工作

2.1　已知协议分析

国际标准化组织公布了开放系统互连参考模型(OSI/RM). OSI/RM 是一种分层的体系结构，参考模型共有 7 层. TCP/IP 作为 Internet 的核心协议，它是个协议簇，包含 FTP、SMTP、TCP、UDP、IP 等多种协议. 以此为例可以得到通常的协议模型，如图 1 所示. 由此可见已知协议都有着公开的文法和规范，通过分析上层协议头中的端口号就能得到下层协议类型. 目前通过 tshark[22]、tcpdump[23]等工具就可以对已知共有协议进行解析分类，但是由于未知协议没有公开的特定端口号以及标准化的协议规范，这些工具都无法进行准确的聚类和分析.

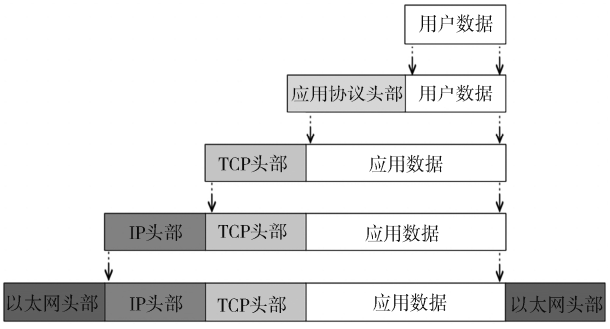

图 1　公有协议模型
Fig. 1　Model of public protocols

2.2　未知协议分析

随着网络规模的不断扩大和应用类型的不断丰富，未知协议的种类也在快速增加. 依据分析对象的不同，可以把未知协议的解析方法分为基于执行轨迹的协议解析方法和基于网络流量的协议解析方法两种.

(1) 基于执行轨迹的协议解析方法面向指令序列，利用动态污点技术跟踪报文数据在通信实体间的解析过程来解析协议. AutoFormat 针对 PI 与 Discover 项目中逆向得到的报文格式为平坦的线性结构，通过构造一个字段树去研究字段之间存在的层次关系(hierarchical field)、并列关系(sequential field)和序列关系(parallel field)，提出了基于指令轨迹的字段结构识别方案. Tupni 沿用 AutoFormat 的思想，利用多条协议信息发现协议格式. 这类方法依赖协议平台，需要执行信息，在使用时的操作性和通用性上存在一定问题.

(2) 基于网络流量的协议解析方法也被称作基于报文的协议逆向技术，分析的对象是从网络上捕获的真实流量. 2004 年 PI(Protocol Informatics)由 Beddoe 启动并发布[16]，在算法中首次引入了生物信息学领域中的渐进多序列对比算法[24]来解决协议逆向工程问题，并根据相同类型分组的统计特征对报文格式进行分析. 之后的 Discover 使用报文序列分析方法实现完整的协议格式提取，模拟了报文解析的层次化过程. 2014 年由 Bossert 等人发起的 Netzob 项目目前已经成为了协议逆向领域的重要开源工具. 孙芳慧[25]针对 PI、Netzob 中的初始聚类方法进行优化并提出了一种将基于协议格式关键字识别的聚类划分方法与基于网络流量特征的聚类归约方法相结合的私有协议格式提取方法. Li 等[2]结合了协议框架的位置信息，设计了一种基于 Jaccard 距离的层次聚类方法. 可见现阶段的协议聚类仍然依赖于人工分析或者文本协议、无法完全应用于包括二进制协议在内的各类协议.

3　未知二进制协议聚类方法

本节介绍进制转换的未知协议聚类解析方法，包括两次多序列比对过程，针对一定数量的协议序列数据，在聚类耗时降低的同时保证了协议种类聚类的准确率.

3.1　方法概述

未知二进制协议聚类方法(Unknown Binary Protocol Clustering，UBPC)采用两次进制转换进行数据预处理. 当前，由于私有未知协议在传输过程中由于安全需要绝大部分不会以明文形式传输，报文数据在转义后也很少以可打印字符的形式出现. 因此 UBPC 没有选取协议报文转义后的字符作为聚类分析的标志性字段进行分析，而是将真实环境下捕获数据包的原始二进制序列数据作为分析的目标，将二进制的协议序列数据进行两次进制转换，即分别把原始协议序列转化成基因序列和蛋白质序列. 对于所有的数据，通过第一次快速聚类方法将所有协议序列预分簇降低聚类时间，再对分

好的每一簇协议序列进行第二次高精度算法聚类提升准确率,整体流程如图 2 所示.

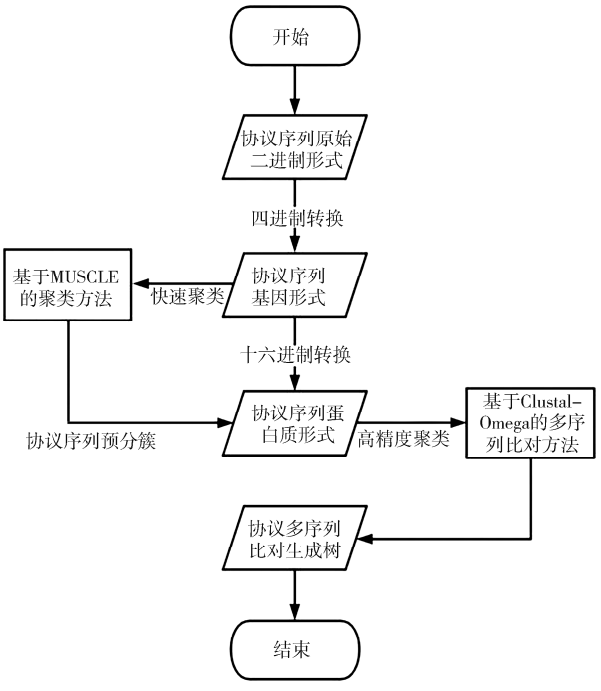

图 2　未知二进制协议聚类方法流程

Fig. 2　Unknown binary protocol clustering method process

3.2　基于 MUSCLE 的聚类方法

此阶段将协议原始的二进制报文序列转化成四进制形式,将其序列长度缩减为之前的 1/2,使用基因序列的 4 种碱基 A,T,G,C 对四进制 0,1,2,3 等 4 种字符进行替换,将其转化成 DNA 序列形式.

首先将其转化成四进制形式,由于四进制形式共包含 0,1,2,3 等 4 种字符,这里在保证语义的基础上选取两个四进制字符代表一个十六进制的组合,因此共有从 00 到 33 十六种组合方式,以滑动窗口大小为 2 从头扫描序列, 统计每种字符组合的个数,而统计后的字符组合记为 k-seed.

需要说明的是,由于这一阶段的目标是产生第一步多重对齐,强调提高聚类速度,所以要求 k-seed 的个数既不能太多,同时还需要保证构造出的 k-seed 距离矩阵能够在一定程度上起到区分协议种类的作用,以保证一定的准确率. 因此将初始序列转化成四进制形式是唯一且必要的. 算法 1 具体流程如下.

算法 1　基于 MUCLE 的聚类方法

Step1　计算距离矩阵 D_1. 计算每对输入序列的 k-seed 距离,采用 Smith-Waterman 算法给出距离矩阵 D_1.

Step2　构造,分割引导树 TREE_1. 矩阵 D 采用非加权成对群算术平均法(Unweighted Pair-Group Method with Arithmetic means, UPGMA)计算子类间的距离逐步将距离最小的子类进行合并,子类 C_i 和 C_j 的距离公式如下,产生二叉树 TREE_1.

$$d_{ij} = \frac{1}{|C_i||C_j|}\sum_{p\in C_i,q\in C_j} D_{pq} \tag{1}$$

Step3　按照 TREE_1 的分支顺序构造渐进对齐. 在每个叶子上,根据输入序列构建配置文件. 树中的节点按前缀顺序访问(子节点在父节点之前). 在每个内部节点,由两个子配置文件构建成对对齐,给出分配给该节点的新配置文件,会在根处生成所有输入序列的多重比对.

在 k-seed 处理后终止算法,选取 Tree_1 作为第一阶段的聚类分簇结果. 由于依照 k-seed 组合的聚类方式经验证,能够在一定程度上作为整体协议序列种类的分类依据[26],在保证了一部分准确率的基础上,在第一次面对大量原始协议序列时,降低了协议数量 N 对聚类时间复杂度的影响.

3.3　基于 Clustal-Omega 的多序列比对方法

由于协议序列中的语义通常是以 4 位二进制数据为单位表示的,因此 UBPC 将原始二进制数据进行十六进制转换,同时将协议报文序列的长度缩减为之前的 1/4,具体时间复杂度分析见 3.4 节. 针对第一阶段的序列簇,使用改进的 mBed 算法得到引导树,算法过程如算法 2 所示. 每次将最近的两个序列进行组合构造引导树,并不断重复,完成了最终的引导树创建,为了提升对齐过程的灵敏度[27],使用了 Soding 的 HHalign 包用于拼接所有渐进对齐.

算法 2　基于 Clustal-Omega 的多序列比对

输入: X 表示第一阶段预分簇的序列集合,P 表示查找有助于表征数据集的潜在种子的算法

输出: 嵌入距离矩阵 D

(1) procedure Clustal O(X)

(2)　　//初始种子选取

(3)　for all $N_i \in X$ do:

(4)　N_i 为簇,令 l_i 为簇中的协议序列个数,遵循 LLR 近似算法,按照默认设置 $t=(\log_2 l_i)^2$ 进行采样,按照恒定步长采样 t 个种子,构造成种子集合 R

(5)　$th=0$//阈值，用于判断种子是否相同
(6)　for all 种子 $s\in R$ do：
(7)　　if $d_{ij}<th$ then
(8)　　丢弃种子
(9)　　else
(10)　　s 添加到 K //K 表示可用种子集合
(11)　end if
(12)　end for
(13)　//潜在种子挖掘
(14)　if $P=$UPO then //UPO 表示单序列查找
(15)　for all 序列 $s\in R$ do
(16)　　令 a 为N_i中使$d(a,s)$最大的序列
(17)　　令 b 为N_i中使$d(b,a)$最大的序列
(18)　　b 添加到K
(19)　end for
(20)　else if UPG then //UPG 表示序列组查找
(21)　　for all $s\in R$ do
(22)　　令 a 为N_i中使$d(a,s)$最大的序列
(23)　　令 b 为N_i中使$d(b,a)+d(b,s)$最大的序列
(24)　　令 c 为N_i中使$d(c,a)+d(c,b)+d(c,s)$最大的序列
(25)　　if 同一序列多次识别 or 数量达上限 then
(26)　　break
(27)　　end if
(28)　　$\{a, b, c, \cdots\}$添加到 K
(29)　　end for
(30)　　end if
(31)　　计算所有序列到种子的 UPGMA 距离 d，使得所有序列都与一个 t 维向量相关联
(32)　　for all 序列 $n\in N_i$ do
(33)　　按照下面公式计算对应的嵌入向量 $F(s)=[d(s,K_1),d(s,K_2),\cdots,d(s,K_t)]$
(34)　　更新 D
(35)　　end for
(36)　end for
(37) end procedure

3.4　效率分析

对于协议报文序列而言，其在数据包中的原始形式如图 3 所示. 假设其中有 N 条 M 种的协议，将每条协议序列记为 Seq，那么，

$$\{Seq_i^k \mid k\in\{0,1,\cdots,M\}, i\in\{0,1,\cdots,N\}\}$$

协议序列长度记为 L.

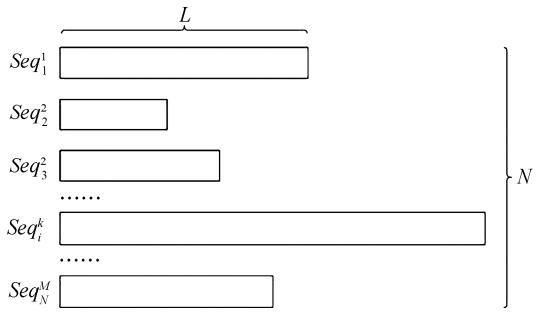

图 3　pcap 数据包序列分布示意图

Fig. 3　Schematic diagram of pcap packet sequence distribution

算法 1 将原始二进制序列转换成四进制形式，序列长度缩短为之前的二分之一，时间复杂度 $O(NL^2)$缩短到四分之一.

算法 2 的时间复杂度为 $O(N\log N)$，因此它的时间开销只与序列数量 N 有关. 算法 1 的预分簇结束后，将大量的协议序列分割，其本身的时间复杂度已经加快到 $O(N\log N)$. 因为分簇的结果没法预知，假设一共预分为 k 簇，那么整体算法时间复杂度计算公式为

$$O(NL^2)+O(N\log N) \tag{2}$$

每一簇协议序列记为$n_i(i=1,2,\cdots,k)$，替换后得时间复杂度计算公式为

$$O(\frac{1}{4}NL^2)+O(\sum_{i=1}^{k} n_i\log n_i) \tag{3}$$

对于每一个$n_i(i=1,2,\cdots,k)$，都可以表示成只有序列个数 N 的表达式，

$$\sum_{i=1}^{k} n_i = N \tag{4}$$

$$n_i=\frac{n_i}{N}\cdot N(i=1,2,\cdots,k) \tag{5}$$

即将每簇的协议序列个数表示为包含一个常数与序列个数 N 的表达式，将式(4) 和式(5)带入式(3)化简后得

$$O\left(\frac{1}{4}NL^2\right)+O\left(N\log\prod_{i=1}^{k} n_i^{\frac{n_i}{N}}\right) \tag{6}$$

根据对数性质，结合式(3)～式(6)可以得出

$$\prod_{i=1}^{k} n_i^{\frac{n_i}{N}} < N \tag{7}$$

在实际情况下，对每一簇可以同时使用算法 2 进行聚类，那么对于不同簇的协议序列$n_i(i=1,2,\cdots,k)$，时间开销更会大大缩短，实际时间复杂度为

$$O(\frac{1}{4}NL^2)+O(\max(n_i\log n_i)) \tag{8}$$

综上所述,由于进制转换缩短了协议序列长度,预分簇减少了聚类时的协议序列数量,因此UBPC与算法单独执行的时间复杂度 $O(NL^2+N^2L+N^3L)$ 和 $O(N\log N)$ 相比,在时间效率上均有一定的提升.

4　算法仿真与实验结果

4.1　实验场景

为了验证上述两次聚类方法能够保证协议聚类高效的同时保持高准确性,本文通过 Wireshark 对真实网络环境下的流量进行捕获,选取其中 5 种已知应用层协议和 3 种私有未知协议作为实验原始数据. 已知协议包括:DNS、HTTP、IGMPv3、LLMNR 和 MDNS. 未知私有协议针对协议的捕获场景分别命名为 TCT、NTE 和 SNA. 为了验证多次聚类处理后对协议识别方法的有效性,将UBPC分别在真实环境下的已知协议、未知协议和已知协议未知协议混合场景下应用. 实验没有单独分析应用层协议,而是使用完整报文协议数据进行验证,使得应用过程更具有普适性.

原始协议序数据需要进行如下预处理,首先分割出完整报文序列;再将报文序列转义出的原始二进制序列进行两种进制转换,即二进制转化为四进制和二进制转化为十六进制;最后将四进制的序列类比成一条基因序列,同理将十六进制的序列类比成一条蛋白质序列,如图 4 所示.

图 4　数据预处理示意图

Fig. 4　Data preprocessing diagram

测试平台选用 MacOS+Python3. 8,硬件使用 MacBook Air 计算机,配以 Apple M1 CPU 和 16 GB 内存.

4.2　已知协议聚类准确性实验

已知协议通常有确定的协议规范,流量中的协议序列也会按照定义的字段结构组成整条协议数据. 因此只包含已知应用层协议的协议序列,通常是链路层、网络层、传输层协议头加应用层的组成方式.

为了验证 UBPC 能够顺利解析协议序列,首先测试其能够对已知协议的种类做到较准确的划分. 实验统计了 UBPC 作用于真实环境下共 981 个协议序列时,测试数据中含有种类个数不同已知协议时的准确率情况,并用其他生物信息学聚类方法的准确率结果作为对比.

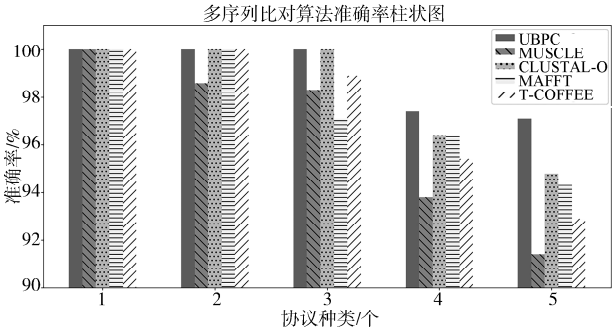

图 5　已知协议聚类准确率实验结果

Fig. 5　Experimental results of known protocol clustering accuracy

由图 5 可以看出,随着协议种类的增加,UBPC 的准确率在各方法对比下一直维持在高水平状态,说明其能够得出较为准确的聚类结果,并且在其他方法中 Clustal-Omega 的准确率保持在非常高的水平,也显示出选用此算法作为高精度聚类算法的依据. 需要说明的是,由于 UBPC 的应用对象是协议的序列数据,对于协议本身是否未知没有判定,因此在协议种类增多后会出现准确率下降的情况,但是仍然可以得到 97%左右的高精度聚类结果.

4.3　未知协议聚类准确率实验

未知协议通常没有统一的格式说明,对于其应用层协议字段没有确定的描述,同样不对协议序列数据进行处理,直接对原始二进制报文开始解析工作.

为了评估 UBPC 对纯未知协议序列数据的解析情况,实验选取了上述三种未知协议共 915 条序列数据,针对包含两种协议和三种协议的混合场景分别进行实验,图 6 分别给出了 5 种方法解析的准确率(%)和所耗时间(s). 其中 x 轴代表时间,y 轴代表解析准确率.

图 6 中较小的图案和较大的图案分别表示两种协议混合场景和三种协议混合场景下的实验结果. 由图 6 可知,在准确率方面,UBPC 和 Clustal-Omega 解析准确率整体来看明显优于其他方法. 且文献[28]指出 Clustal-Omega 在大量数据下的处理速度非常优秀. 我们通过进制的转化将二进制

转化成四进制和十六进制，将协议序列长度分别缩减为之前的 1/2 和 1/4，降低了序列长度过长所带来的时间开销. 因此在时间已经非常短的情况下又进一步提升了解析效率. 在时间方面，UBPC 通过引入更精准的渐进比对算法，使得时间开销逼近甚至赶超目前最快的方法 MAFFT[28].

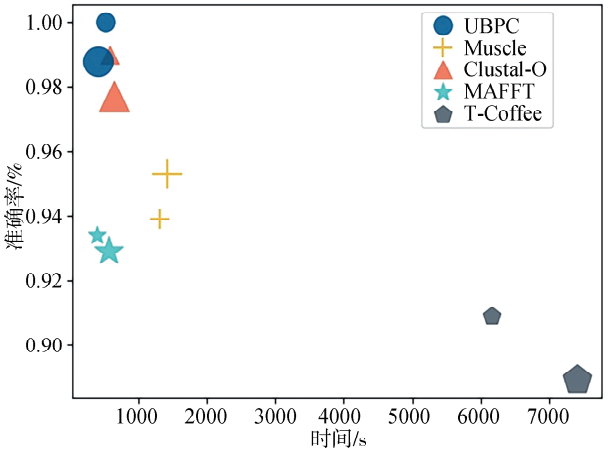

图 6　未知协议聚类准确率和效率实验结果

Fig. 6　Unknown protocol clustering accuracy and efficiency experimental results

4.4　真实场景已知未知协议混合效率实验

真实场景下的流量数据通常是已知和未知协议混合的情况. 为了观察 UBPC 在这种场景下的表现，选取的数据包部分包括 DNS、HTTP、IGMPv3、LLMNR 和 MDNS 5 种已知协议及 TCT、NTE 和 SNA 3 种未知协议共 8 种 1896 条协议序列，在相同硬件条件环境下进行实验.

表 1　真实环境聚类准确率和效率实验结果

Tab. 1　Clustering accuracy and efficiency experimental results in real environments

方法	准确率/%	时间/s
UBPC	99.7	1289
Muscle	98.5	3317
Clustal-O	99.4	8102
MAFFT	96.3	1358
T-Coffee	75.3	36 862

由表 1 可以看出，UBPC 可以很好地对真实环境中的已知和未知协议序列进行较为准确的区分，在与单独使用 Cluster-Omega 的准确率基本持平的情况下，UBPC 的时间开销有了明显的降低. 另外，对比效率较高的 MAFFT 方法，在时间开销基本持平的同时，UBPC 的准确率得到了明显的提升. 因此与其他方法相比，UBPC 方法在时间开销和准确率的表现上都有着不俗的表现.

5　结　论

未知协议逆向工程的研究取得了一定的成果，而面向二进制协议报文序列的未知协议聚类作为提高协议逆向准确率和效率的重要步骤也逐渐成为了研究的热点问题. 本文摆脱了大量研究和开源工具对于文本类协议、协议特征和人工分析的依赖，从二进制序列本身入手，将其进行两次进制转换，结合基因和蛋白质的多序列比对方法，通过第一次的高效聚类方法降低时间开销和第二次高精度聚类方法提升准确率，对已知协议、未知协议和真实环境下的混合协议序列进行聚类分析，并与其他序列聚类方法进行对比，结果证明 UBPC 准确率和效率都更加出色.

参考文献：

[1]　党建网微平台. 习近平谈网络安全[EB/OL]. (2021-09-27)[2021-09-28]. https://mp.weixin.qq.com/s/6n84MHjRU-4m0qWhTR4-aQ.

[2]　Li Y, Bai L, Zhang M, *et al*. Network protocol reverse parsing based on bit stream [C]// Proceedings of the 2021 8th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2021 7th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom). Washington, DC, USA: IEEE, 2021.

[3]　张光华，石晓朦. 面向应用层协议的自动化模糊测试方案[J]. 微电子学与计算机，2018，35：99.

[4]　秦中元，陆凯，张群芳，等. 一种二进制私有协议字段格式划分方法[J]. 小型微型计算机系统，2019，40：2318.

[5]　Google group. Samba documents provider [EB/OL]. (2017-10-21)[2021-09-27]. https://github.com/google/samba-documents-provider.

[6]　Yun X, Wang Y, Zhang Y, *et al*. A semantics-aware approach to the automated network protocol identification[J]. IEEE ACM T Network, 2015, 24: 583.

[7]　周帅，王绍杰. 私有工控协议分类方法研究[J]. 信息技术与网络安全，2021，40：19.

[8]　彭博一，张钊，蒋鸿宇，等. 一种基于改进自编码器的二进制协议聚类方法[J]. 太赫兹科学与电子信息学报，2021，19：712.

[9]　Cui W, Kannan J, Wang H J. Discoverer: automat-

ic protocol reverse engineering from network Traces [C]//Proceedings of the 16th USENIX Security Symposium. Boston, MA(US): USENIX, 2007.

[10] 潘璠, 洪征, 杜有翔, 等. 基于递归聚类的报文结构提取方法[J]. 四川大学学报: 工程科学版, 2012, 44: 137.

[11] Bossert G, Guihéry F, Hiet G. Towards automated protocol reverse engineering using semantic information [C]//Proceedings of the 9th ACM symposium on Information, Computer And Communications Security. Kyoto, Japan: ACM, 2014.

[12] Hiet G, Tong V V T, Me L, *et al*. Policy-based intrusion detection in web applications by monitoring java information flows [C] // Proceedings of the 2008 Third International Conference on Risks and Security of Internet and Systems. Tozeur, Tunisia: IEEE, 2008.

[13] 王晓晨, 沈晶, 刘海波, 等. 自动协议逆向工程研究综述[J]. 计算机应用研究, 2020, 37, 2561.

[14] Yan X, Li Q, Tao S. A clustering algorithm for binary protocol data frames based on principal component analysis and density peaks clustering [C] // Proceedings of the 2017 IEEE 17th International Conference on Communication Technology (ICCT). Chengdu, China: IEEE, 2017.

[15] 张路煜, 廖鹏, 赵俊峰, 等. 基于卷积神经网络的未知协议识别方法[J]. 微电子学与计算机, 2018, 35: 106.

[16] Tao S, Yu H, Li Q. Bit-oriented format extraction approach for automatic binary protocol reverse engineering [J]. IET Commun, 2016, 10: 709.

[17] Lin Z, Jiang X, Xu D, *et al*. Automatic protocol format reverse engineering through context-Aware monitored execution [C]. San Diego, California, USA: ISOC, 2008.

[18] Cui W, Peinado M, Chen K, *et al*. Tupni: Automatic reverse engineering of input formats [C]// Proceedings of the 15th ACM conference on Computer and Communications Security. Alexandria, Virginia, USA: ACM, 2008.

[19] Wang Y, Li X, Meng J, *et al*. Biprominer: Automatic mining of binary protocol features[C]//Proceedings of the 2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies. Gwangju, Korea: IEEE, 2011.

[20] Wang Y, Yun X, Shafiq M Z, *et al*. A semantics aware approach to automated reverse engineering unknown protocols [C]// Proceedings of the 2012 20th IEEE International Conference on Network Protocols (ICNP). Austin, TX, USA: IEEE, 2012.

[21] Zhang Z, Zhang Z, Lee P P C, *et al*. Toward unsupervised protocol feature word extraction [J]. IEEE J Sel Area Comm, 2014, 32: 1894.

[22] 林帅, 王俊峰. 基于 Tshark 的可扩展报文解析功能的实现[J]. 网络安全技术与应用, 2016, 4: 49.

[23] Bonelli N, Giordano S, Procissi G. Enabling packet fan-out in the libpcap library for parallel traffic processing [C] // Proceedings of the 2017 Network Traffic Measurement and Analysis Conference (TMA). Dublin, Ireland: IEEE, 2017.

[24] 王慧锋, 段磊, 左劼, 等. 免预设间隔约束的对比序列模式高效挖掘[J]. 计算机学报, 2016, 39: 1979.

[25] 孙芳慧. 基于 Net-Trace 的未知协议格式逆向技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2015.

[26] Edgar R C. MUSCLE: multiple sequence alignment with high accuracy and high throughput [J]. Nucleic Acids Res, 2004, 32: 1792.

[27] Reddy B, Fields R. Multiple sequence alignment algorithms in bioinformatics [M]//Smart Trends in Computing and Communications. Singapore: Springer, 2022.

[28] Sievers F, Wilm A, Dineen D, *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega [J]. Mol Syst Biol, 2011, 7: 539.

引用本文格式:

中　文: 丛培鑫, 李晓慧, 王俊峰. 基于生物信息的未知二进制协议聚类方法[J]. 四川大学学报: 自然科学版, 2022, 59: 032004.

英　文: Cong P X, Li X H, Wang J F. Unknown binary protocol clustering method based on biological information [J]. J Sichuan Univ: Nat Sci Ed, 2022, 59: 032004.