计算机科学
COMPUTER SCIENCE

基于概率模型的二进制协议字段划分方法

杨资集，潘雁，祝跃飞，李小伟

引用本文

杨资集，潘雁，祝跃飞，李小伟. 基于概率模型的二进制协议字段划分方法[J]. 计算机科学, 2022, 49(10): 319-326.

YANG Zi-ji, PAN Yan, ZHU Yue-fei, LI Xiao-wei. Field Segmentation of Binary Protocol Based on Probability Model[J]. Computer Science, 2022, 49(10): 319-326.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210800268

基于概率模型的二进制协议字段划分方法

杨资集　潘　雁　祝跃飞　李小伟
战略支援部队信息工程大学　郑州 450001
数字工程与先进计算国家重点实验室　郑州 450001
(zijiyang@yeah.net)

摘　要　字段划分是协议格式推断的基础,协议格式推断的后续步骤,如报文结构识别、字段语义推断和字段取值约束判定,高度依赖于字段划分质量。二进制协议缺少字符编码和定界符,字段长度取值灵活,值域变化丰富,因此字段划分难度较大。针对相关研究存在的特征构造维度单一和判决规则简单等问题,提出了一种基于概率模型的二进制协议字段划分方法。以二进制协议报文为研究对象,从报文内在结构、报文间取值变化等维度构造字段边界约束关系,然后用概率的方式将各种约束组合在一起,利用因子图模型计算各个位置成为边界的概率,从中得出最有可能的字段边界。实验结果表明,相比传统方法,所提方法在二进制协议字段边界识别中精准度更高、鲁棒性更强。

关键词:字段划分;因子图;概率模型;协议逆向
中图法分类号　TP393

Field Segmentation of Binary Protocol Based on Probability Model

YANG Zi-ji,PAN Yan,ZHU Yue-fei and LI Xiao-wei
Strategic Support Force Information Engineering University,Zhengzhou 450001,China
State Key Laboratory of Mathematical Engineering and Advanced Computing,Zhengzhou 450001,China

Abstract　Field segmentation is the basis of protocol format inference. The subsequent steps of protocol format inference,such as message structure identification,field semantic inference and field value constraint inference,highly depend on the quality of field segmentation. Field segmentation of binary protocol is a big challenge because of the lack of character coding and delimitation,the flexibility of field length and the expansiveness of field range. To improve feature construction and decision rules,this paper proposes a novel binary protocol field segmentation method based on probability model. First,it constructs the field boundary constraint relationship of binary protocol messages from the internal structure of message and the value change between messages. Then,it combines various constraints in the way of probability,calculating the probability of each position becoming the boundary by factor graph model. Finally,the most likely field boundaries are obtained from probability. Experiments show that the proposed method can achieve more accurate and robust results than the traditional methods in binary protocol field segmentation.

Keywords　Field segmentation,Factor graph,Probability model,Protocol reverse

1　引言

网络协议是计算机网络中通信节点进行数据交换时必须遵守的一组规则,这些规则由语法、语义和同步3个要素组成。随着互联网的快速发展,诞生了许多公开的、标准化的网络协议。同时,出于经济利益、隐私保护、秘密通信等考虑,网络中还存在很多未知协议。根据安全公司Sophos在2018年发起的一项针对全球中小企业的网络安全调查显示,参与单位的网络流量中无法被识别的流量平均占比达到45%[1]。未知协议的出现,不仅增加了网络管理的难度,而且使得网络攻击者可以隐蔽地进行攻击活动,给人们带来了潜在的网络安全隐患。

为了应对未知协议带来的挑战,越来越多的研究采用协议逆向的手段对未知协议进行分析。协议逆向工程指在不依赖先验的协议描述信息的情况下,通过对协议实体的网络行为、系统行为和指令执行流程进行监控和分析,实现协议格式分析、协议语义推断、协议状态机建模的过程[2]。协议逆向分析有基于网络流量分析和基于执行轨迹分析两种方法[3]。前者仅以报文数据作为分析对象,无法精准解析协议的格式,但该方法不需要对任何通信实体进行控制,具有分析速度快、容易实现的优点。本文采用的是基于网络流量分析的方法。

以应用层协议为例,基于网络流量的协议格式分析流程

到稿日期:2021-08-30　返修日期:2021-12-03
基金项目:国家重点研发计划(2019QY1300)
This work was supported by the National Key R&D Program of China(2019QY1300).
通信作者:祝跃飞(yfzhu17@sina.com)

如图 1 所示。协议格式分析主要以协议数据单元为研究对象，即目标协议所处层次传递的数据单元，数据包、报文段和报文分别是网络层、运输层和应用层的数据传递单元。以二进制协议报文为研究对象，本文的研究内容聚焦于协议字段划分。

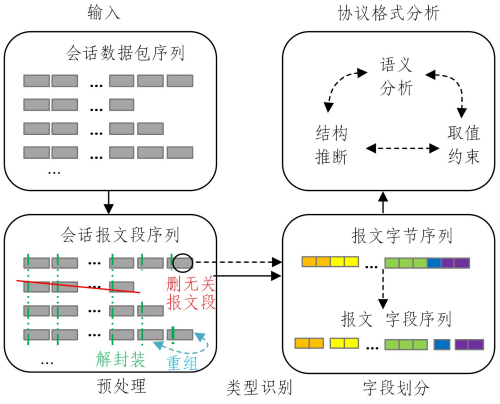

图 1 基于网络流量的协议格式分析

Fig. 1 Protocol format inference based on network traffic

二进制协议缺少字符编码和定界符，字段长度取值灵活，值域变化丰富，因此字段划分难度较大。当前基于网络流量分析的二进制协议字段划分技术主要存在以下两个方面的问题：

(1)特征构造方面，当前的分析方法主要是基于自然语言处理算法和生物信息学算法，这两种方法通过关键字和对齐的序列来构造边界特征。然而，许多协议没有使用自然语言处理可以识别的关键字，而精确多序列比对由于其指数复杂性，无法对大量的报文进行分析。少量基于报文结构来构造特征的方法则存在误判率高、边界偏移等问题。目前还缺少能借鉴不同技术从多维度构造特征且适用范围广的方法。

(2)边界判决方面，当前工作普遍使用观察后设定的规则或简单的启发式方法来确定边界位置，但是边界特征的构造本身与真实边界分布存在偏差，具有内在的不确定性，任意字节序列都可能满足构造的特征分布，因此简单的判决准则会导致结果的冗余和偏差，需要建立合理评估方法来对字段边界进行准确判决。

针对当前方法存在的特征构造维度单一的问题，本文在对齐的基础上从报文内在结构(字节相似性)、报文间取值变化(信息熵、互信息熵、字段类型)等多维度构造字段边界约束关系，所有约束都是对字段边界的一种提示，不同约束的组合能够更准确地反映字段边界的分布。

针对当前方法存在的边界判决规则简单的问题，受先前工作[4-6]在协议格式逆向分析中应用概率推理的启发，本文以概率的方式将各种约束组合在一起，充分利用多种约束完成对字段边界的最优推断。具体而言，通过观察并利用先验知识为每个提示边界分布的约束关系计算先验概率，然后使用因子图模型聚合这些提示并计算各个位置成为边界的概率，得到候选字段边界，最后结合启发式规则进行修正。

本节讨论了基于网络流量分析的二进制协议字段划分的背景和存在的问题，并提出了解决方法；第 2 节总结了国内外相关工作；第 3 节介绍了本文构造的 5 种约束关系和基于因子图模型的字段边界判决方法；第 4 节展示了实验结果并进行了分析；最后总结全文并展望未来。

2 相关工作

对于基于网络流量分析的协议字段划分，目前国内外已有许多的相关研究工作。根据可处理对象的不同，协议字段划分技术可分为文本协议字段划分技术和二进制协议字段划分技术，前者仅能分析文本协议，后者专注于二进制协议，但也可用于文本协议。

在早期阶段，相关研究工作以文本协议字段划分技术为主。Beddoe[7]于 2004 年提出了 PI 项目(Protocol Informatics Project)，该项目使用多序列比对算法对报文字节序列进行对齐，根据对齐结果来确定固定与可变字段域。2006 年，Cui 等[8]提出的 RolePlayer 首先使用启发式方法和先验知识来确定感兴趣的字段，再通过序列比对发现可变字段域。针对基于字节的序列比对算法识别变长字段效果差的缺点，2007 年，Cui 等[9]提出了 Discoverer。该方法设计了一种基于类型的序列比对方法，首先根据 ASCII 编码划分文本段和二进制段，然后基于划分段的类型序列进行比对来确定报文格式和类型。2010 年，Krueger 等[10]提出了一种基于字符编码集的 t-test 假设检验方法，用于分析固定与可变字段域。2012 年，Pan[11]等提出了一种基于递归聚类的报文结构提取方法，在初步划段和聚类的基础上，基于多序列比对和语义推断进行字段边界划分和报文类别判定。

随着研究的深入，对二进制协议字段划分技术的需求越来越大。在 2011 年和 2012 年，Wang 等[12-13]针对 Discoverer 方案的字符边界缺陷，提出了使用 n-gram 对关键字进行建模的方法，分别利用概率转移矩阵和隐含狄利克雷分配模型对关键字进行识别。Li 等[14]于 2013 年提出了一种利用隐半马尔可夫模型对未知应用层协议进行最佳分段的方法。2014 年，Bossert 等[15]提出了 Netzob，该工具在字段划分过程中使用了上下文信息及其语义作为关键字参数。同年，Zhang 等[16]在提出的方法 ProWord 中首次引入投票专家算法和信息熵进行字段划分。Bermudez 等[17-18]于 2016 年提出的 FieldHunter 分别使用 n-gram 和识别分隔符的方法对二进制和文本协议进行分段。2018 年，Kleber 等[19]提出了一种利用网络报文中取值变化的典型模式来推断字段边界的方法，该方法使用报文连续字节的相似性来感知消息内在结构。2019 年，Sun 等[20]通过引入和优化信息论中的统计量，提出了一种针对固定字段长度的二进制协议的字段划分方法 ProSeg。2020 年，Jiang 等[21]提出了一种利用字段邻接信息来识别字段边界的方法 ABInfer，Wang 等[22]提出了一种针对工业协议的自动化逆向工具 IPART，该工具结合全局投票专家与重定位方法来推断字段边界。Ye 等[4]于 2021 年提出了 NETPLIER，该方法在序列比对基础上以字节为单位通过启发式规则识别固定、动态和变长字节，再将连续相同类型的字节合并为字段。同年，Liu 等[23]提出了 Entry Distance Cluster 算法，用于递归合并相邻字段并推断字段边界。

相比上述相关工作，本文的创新点在于充分利用近年来

提出的多种字段边界特征，并通过概率模型组合所有特征，完成对字段边界的综合推断。本文方法的总体流程如图 2 所示。

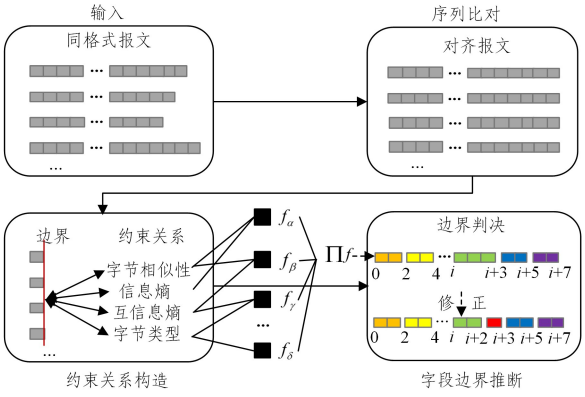

图 2 总体流程

Fig. 2 Overall process

3 方法设计

3.1 总体流程

首先对流量数据进行预处理，获取同格式报文作为输入，然后使用多序列比对算法对同格式报文进行对齐，在对齐报文上进行多维度的约束关系构造，最后基于因子图模型判决字段边界并利用启发式规则进行修正。

3.2 输入与序列比对

在协议格式已知的情况下，获取同格式的报文作为输入较为简单，而对于未知协议，可以将所有报文一起输入获得公共的格式，也可以简单地把报文根据方向分为请求和响应两个类别。更精确的做法是通过聚类把报文分成几类格式相似的报文集。报文聚类指利用报文之间的相似度对其类型进行分组，常见的方法有机器学习的聚类算法和自定义的特定分类方法。这部分不是本文研究的重点，为便于评估，我们使用已知协议来模拟未知协议，获取同格式的报文作为输入。

序列比对算法的理论基础是进化学，最初用于生物学，目的是通过比对 DNA、RNA 和蛋白质序列来识别相似区域，对于揭示生物序列中的结构、功能和进化信息意义重大[24]。网络协议本质上是一种语言，序列比对可利用协议报文之间的进化相似性对其格式进行分析研究。

本文使用序列比对算法对同格式的报文进行对齐，而后在对齐的报文上构造约束关系，具体使用的工具是 MAFFT 版本 7[25]。MAFFT 是一种基于快速傅里叶变换的多序列比对方法，该方法可以快速检测同源片段，是公认的多序列比对最佳实现方案之一。

3.3 约束关系构造

当前用于构建映射协议字段边界的拟合特征的方法主要可以分为基于序列比对、基于图形概括和基于代数分析 3 种[26]。为了对同格式报文中的字段边界特征进行深入分析，本文结合基于序列比对的方法和基于代数分析的方法，在对报文进行对齐的基础上，设计了 5 种基于代数分析的方法来构造字段边界约束关系，以拟合字段边界特征。

在构建约束关系之前，有必要先给出报文字段相关的形式化描述。我们假设报文字段都由字节组成，也就是说，字段边界的位置都以字节为单位，这个单位在需要时可以用半字节或比特来代替。如图 3 所示，对于同一格式的对齐报文集 $M=\{msg_1,msg_2,\cdots,msg_n\}$，每条报文可以表示为 $msg=(byte_1,byte_2,\cdots,byte_m)$。其中 $byte_i$ 取值介于 0x00 和 0xff 之间，表示报文的第 i 字节；n 和 m 分别表示报文集中报文数量和对齐后报文的字节数。字节可进一步表示为二进制形式 $byte=(bit_1,bit_2,\cdots,bit_8)$，其中 $bit\in\{0,1\}$。除了水平方向，报文集还可以表示为一连串竖向字节序列 $V=(v_1,v_2,\cdots,v_m)$。本文的目标是判断报文集 M 的哪些字节偏移是其字段边界，即得到字段边界标识 $L=(l_1,l_2,\cdots,l_m)$，其中 $l_i\in\{0,1\}$，$l_i=1$ 表示 i 字节偏移位置是字段边界，否则不是。为便于表达，令 $B(i)$ 表示 $l_i=1$ 这一事件，$B(i)$ 具有不确定性，其概率表现为一个随机变量。

	v_1	v_2	…	v_m
msg_1	byte	byte	…	byte
msg_2	byte	byte	…	byte
⋮	…	…	…	…
msg_n	byte	byte	…	byte

图 3 对齐报文集

Fig. 3 Aligned message set

3.3.1 字节相似性约束

每个报文本身都包含其报文结构的特定提示，例如，二进制协议通常使用通用数据类型长度作为字段长度，而字段内容并不能均匀地填充这种固定长度字段，这种不均匀填充会以特定的方差分布等特征提示字段边界的位置。以 4 字节整数字段为例，当字段值为 2 069 时，其十六进制表示为 0x00000815，我们容易根据连续的 0 判断字段起始位置。借鉴 NEMESYS 的方法，本节以报文中连续字节的相似性特征来构造报文内在结构与 $B(i)$ 的约束关系。

定义字节相似率为报文中连续两字节对应位置取值相同的比特数占一个字节的比率，令 bit_k^i 表示 $byte_i$ 二进制形式的第 k 位，则 i 字节偏移位置字节相似率定义如式(1)所示：

$$BC_i=\frac{|\{bit_k^i=bit_k^{i+1},1\leqslant k\leqslant 8\}|}{8} \tag{1}$$

定义字节相似率增量为相邻字节相似率的差，定义字节相似率增量的变化率为相邻字节相似率增量的差，如式(2)和式(3)所示：

$$\Delta BC_i=BC_i-BC_{i-1} \tag{2}$$

$$\Delta_{\Delta BC}^i=\Delta BC_i-\Delta BC_{i-1} \tag{3}$$

根据相关工作[19]，$\Delta_{\Delta BC}^i$ 与 $B(i)$ 存在约束关系，如果 i 字节偏移处为字段边界，那么 msg 的 i 字节偏移处的 $\Delta_{\Delta BC}^i$ 往往为所处 ΔBC 上升区间最大值。定义 msg 的 i 字节偏移处的 $\Delta_{\Delta BC}^i$ 是所处 ΔBC 上升区间最大值为 $S(i,msg)$，$S(i,msg)$ 和 $B(i)$ 存在两种约束关系：观察约束和推断约束。

观察约束指 $S(i,msg)$ 本身是否成立，通过计算报文上 i 字节偏移处的 $\Delta_{\Delta BC}^i$ 是否为所处 ΔBC 上升区间最大值，可以得到 $S(i,msg)$ 的先验观察概率 p_m。式(4)给出了单条报文 i 字节偏移处概率提示 p_m 的赋值方法，若 i 字节偏移处的 $\Delta_{\Delta BC}^i$ 为所处 ΔBC 上升区间最大值，则将 p_m 赋值为 0.9，否则赋值为 0.1。如式(5)所示，对每条报文的概率提示取平均值可得到报文集 i 字节偏移处 $S(i,M)$ 的观察概率 p_s，其中，$S(i,M)$ 定义为所有报文上 i 字节偏移处的 $\Delta_{\Delta BC}^i$ 均为

所处 ΔBC 上升区间最大值。

$$p_m=\begin{cases}0.9, & \Delta^{i-1}_{\Delta BC}<\Delta^{i}_{\Delta BC} \text{ and } 0<\Delta^{i+1}_{\Delta BC}<\Delta^{i}_{\Delta BC}\\ 0.1, & \text{others}\end{cases} \tag{4}$$

$$p_s=\frac{\sum_{k=1}^{n}p_m^k}{n} \tag{5}$$

推断约束指 $S(i,M)$和 $B(i)$之间的推断关系，比如 $B(i)$成立，能否推断 $S(i,M)$成立，可以用 $p_{s\rightarrow}$表示推断成立的概率；同理，$p_{s\leftarrow}$表示在 $S(i,M)$成立的情况下，$B(i)$成立的概率。$p_{s\rightarrow}$和 $p_{s\leftarrow}$通常利用先验知识设定。

3.3.2 信息熵约束

对于大多数协议，每个特定字段都携带相应的信息，具备相应的功能，因此不同的字段具有不同的数据分布。由于每个字段所承载的信息量与相邻字段之间存在差异，信息论领域的度量方法信息熵可以很好地对字段边界进行度量。

作为信息含量和多样性指数的度量，信息熵的值可以用来表示不同字段的信息量。报文集中竖向字节序列 v_i 的信息熵 E_i 可以指示该序列携带的数据信息量，其定义如式(6)所示。对于某些特定字段，其字段内部的数据通常携带类似的信息量，因此具有接近的信息熵且其信息熵倾向于随着字节偏移的增加而增加，而信息量的下降通常意味着字段的结束[20]。定义 $E(i,V)$为 v_i 的信息熵 E_i 是 V 中极大值，则$E(i,V)$和 $B(i)$同样存在两种约束关系。观察概率 p_e 的计算式如式(7)所示，推断概率 $p_{e\rightarrow}$和 $p_{e\leftarrow}$可通过先验知识确定。

$$E_i=-\sum_{c\in v_i}P(c)\log P(c) \tag{6}$$

$$p_e=\begin{cases}0.9, & E_{i-1}<E_i \text{ and } E_{i+1}<E_i\\ 0.1, & \text{others}\end{cases} \tag{7}$$

3.3.3 互信息熵约束

除了信息熵，根据信息理论中的定义，互信息 $U_{i,i+1}$可以表示 V 中两个相邻字节序列 v_i 和 v_{i+1}的统计相关程度，其定义如式(8)所示。对于报文中的协议字段，两个连续字节的互信息值可以提示它们之间的统计相关性。对于字段内具有强相关性的连续数据，对应的互信息值通常较高；对于字段间的数据，取值则较低[20]。定义 $U1(i,V)$为 v_i 和 v_{i+1}的互信息熵低于 0.05，$U1(i,V)$和 $B(i)$的两种约束关系对应的概率为 p_{u1}，$p_{u1\rightarrow}$和 $p_{u1\leftarrow}$，其中 p_{u1}的计算式如式(9)所示。此外，在同一的字段中，互信息熵具有接近且相对较高的值，在边界处通常存在极小值[20]。定义 $U2(i,V)$为 v_i 和 v_{i+1}的互信息熵为极小值，$U2(i,V)$和 $B(i)$的两种约束关系对应的概率为 p_{u2}，$p_{u2\rightarrow}$和 $p_{u2\leftarrow}$，其中 p_{u2}的计算式如式(10)所示：

$$U_{i,i+1}=-\sum_{c\in v_i,d\in v_{i+1}}P(cd)\log\frac{P(cd)}{P(c)P(d)} \tag{8}$$

$$p_{u1}=\begin{cases}0.9, & U_{i,i+1}<0.05\\ 0.1, & \text{others}\end{cases} \tag{9}$$

$$p_{u2}=\begin{cases}0.9, & U_{i,i+1}<U_{i-1,i} \text{ and } U_{i,i+1}<U_{i+1,i+2}\\ 0.1, & \text{others}\end{cases} \tag{10}$$

3.3.4 字段类型约束

在序列比对的基础上，可以在垂直方向对报文集每个字节的取值进行比较并判断竖向字节序列的类型，类型通常分为固定值、动态值和变长值 3 类。如果某字节位置出现对齐填充符，则将该字节标记为变长值；在没出现对齐填充符的情况下，如果所有报文数据都具有相同的值，则将该字节标记为固定值，否则为动态值。随后将连续的静态值与变长值合并形成静态字段和变长字段，使用以字节为单位的动态值直接作为动态字段。通过取值变化划分的字段与真实字段有一定的吻合度[4]，定义 $T(i,V)$为 v_i 和 v_{i+1}的字段类型不同，令 tag_i 表示 v_i 所属的字段类型，$T(i,V)$和 $B(i)$的两种约束关系对应的概率为 p_t，$p_{t\rightarrow}$和 $p_{t\leftarrow}$，其中 p_t 的计算式如式(11)所示：

$$p_t=\begin{cases}0.9, & tag_i\neq tag_{i+1}\\ 0.1, & \text{others}\end{cases} \tag{11}$$

3.4 字段边界推断

本文方法的一个关键步骤是将字段边界识别中的不确定性建模为观察值和随机变量的联合分布，观察值即为前面提到的几种约束关系，变量则表示一个候选位置是否是报文的字段边界。这一节首先展示了使用概率模型进行建模和使用因子图进行概率计算的过程，然后设置阈值进行字段边界推断并结合启发式规则进行修正。

首先，使用概率函数描述各个约束关系，为简化表达，令 b 代表 $B(i)$，x_1-x_5代表 5 种观察，以字节相似性约束为例，令 x_1表示 $S(i,M)$，则其观察约束和两个推断约束如式(12)—式(14)所示：

$$f(x_1)=\begin{cases}p_s, & x_1 \text{ is true}\\ 1-p_s, & x_1 \text{ is not true}\end{cases} \tag{12}$$

$$f(b,x_1)=\begin{cases}p_{s\rightarrow}, & b\rightarrow x_1 \text{ is true}\\ 1-p_{s\rightarrow}, & b\rightarrow x_1 \text{ is not true}\end{cases} \tag{13}$$

$$f(x_1,b)=\begin{cases}p_{s\leftarrow}, & x_1\rightarrow b \text{ is true}\\ 1-p_{s\leftarrow}, & x_1\rightarrow b \text{ is not true}\end{cases} \tag{14}$$

然后，令 f_k表示各个概率函数，所有变量的联合概率函数可以表示为概率函数取值的乘积除以概率函数所有可能取值乘积的和，如式(15)所示：

$$p(b,x_1,x_2,x_3,x_4,x_5)=\frac{\prod_{k=1}^{15}f_k}{\sum_{b,x_1,x_2,x_3,x_4,x_5}\prod_{k=1}^{15}f_k} \tag{15}$$

最后，i 字节偏移位置为边界的概率可表示为 b 的边缘概率，如式(16)所示：

$$p(b)=\sum_{x_1,x_2,x_3,x_4,x_5}p(b,x_1,x_2,x_3,x_4,x_5) \tag{16}$$

具体计算时，我们用因子图来表示所有的概率函数并进行高效概率计算。作为一种概率图模型，因子图除了能解决数学概率问题外，还被广泛应用于信号处理、系统生物学和系统动力学等领域[27-30]。因子图是具有因子节点和可变节点的二分图，因子节点代表概率函数，变量节点表示在概率函数中使用到的变量，当变量节点和因子节点相关时，它们中间会产生一条连线，本文使用的字段划分因子图模型如图 4 所示。因子图能通过信念传播算法高效地求解各个变量的边缘分布，具体实现可参考 Ankan 等的研究[31]。

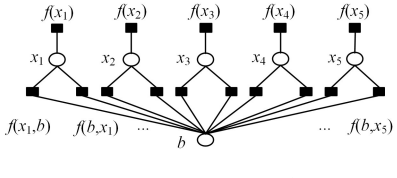

图 4 字段划分因子图模型

Fig. 4 Factor graph model of field segmentation

得到所有字节偏移位置的概率后，我们通过几个约束关系提示不一致的位置来确定阈值 *threshold*（具体方法见下节），所有概率大于或等于 *threshold* 的字节偏移位置被判定为字段边界。*threshold* 可以根据需要上下调节，比如在输入报文集存在不同格式报文时，可以调高 *threshold* 以获得更保守的结果。对于初步得到的字段边界，通过观察发现：1）字节相似性约束无法判别首字节偏移位置，而字节类型约束倾向于把首字节划分为边界；2）字段长度一般不会取除 1 以外的奇数，但推断得到的结果中，相邻边界存在差值大于 1 的奇数。

针对上述问题，我们通过以下两个启发式规则对结果进行修正：1）首字节偏移位置仅考虑信息熵和互信息熵约束关系；2）拆分长度为大于 1 的奇数的字段，具体而言，由低字节到高字节，把长度为 3 的拆为 1 和 2，长度为 5 的拆为 1，2，2，长度为 7 的拆为 1，2，4，长度为 9 的拆为 1，4，4，其余情况不予考虑。

4 实验分析

本文实验使用 python3.7 进行编程，计算机配置双核 i7CPU，16 GB 内存，运行 64 位的 ubuntu18.04 操作系统。为了尽可能模拟未知二进制协议的特性，我们选取了 4 种工控协议（s7comm，modbus，dnp3 和 Ethernet/IP）数据作为实验数据，每种协议的数据集都有 100 条报文和 1 000 条报文两组。上述协议都拥有不只一种报文格式，对于 s7comm 和 Ethernet/IP，我们仅测试所有报文的公共格式；对于 modbus 和 dnp3，我们区分了不同格式的报文，取平均值后进行结果展示。

4.1 评估指标

参考 Ye 等[4]和 Bossert 等[15]的工作，本文选取了 4 个实验结果评估指标，分别为准确率、召回率、边界 F1 值和字段 F1 值。准确率定义为真实边界在推断边界集合中所占的比率，召回率定义为推断正确的边界在真实边界集合中所占的比率，边界 F1 值定义为准确率和召回率的调和平均，是一种综合性能度量。除了从边界的角度衡量实验结果，考虑到实际使用中要进一步组合边界来确定字段，本文还选取了字段 F1 值作为评价指标，字段 F1 值定义为字段准确率和字段召回率的调和平均。准确率主要反映边界推断的准确性，对于推断结果中出现的多余划分容忍度较低，而对于少划分的情况容忍度较高；召回率主要反映边界推断的完整性，对于推断结果中出现的少划分的情况容忍度较低，而对于多余划分容忍度较高；边界 F1 值和字段 F1 值主要反映边界和字段推断的综合性能，后者对于边界推断结果中连续的边界准确性要求较高。

假设真实的协议字段边界集合为 $O_{\text{true}}=\{t_1,t_2,\cdots,t_k\}$，推断的协议字段边界集合为 $O_{\text{infer}}=\{i_1,i_2,\cdots,i_l\}$，相邻的边界表示一个字段的范围，因此 $F_{\text{true}}=\{(0,t_1),(t_1+1,t_2),\cdots,(t_{k-1}+1,t_k)\}$可以表示真实协议字段集合，推断字段集合则可以表示为 $F_{\text{infer}}=\{(0,i_1),(i_1+1,i_2),\cdots,(i_{l-1}+1,i_l)\}$。

根据上述定义，准确率 P、覆盖率 R、边界 $F1$ 值 $F1$ 和字段 $F1$ 值 $fld\text{-}F1$ 这 4 种评估指标对应的计算式如式（17）－式（20）所示：

$$P=\frac{|\{o|o\in O_{\text{true}}\bigcap O_{\text{infer}}\}|}{|O_{\text{infer}}|}\times 100\% \tag{17}$$

$$R=\frac{|\{o|o\in O_{\text{true}}\bigcap O_{\text{infer}}\}|}{|O_{\text{true}}|}\times 100\% \tag{18}$$

$$F1=\frac{2|\{o|o\in O_{\text{true}}\bigcap O_{\text{infer}}\}|}{|O_{\text{true}}|+|O_{\text{infer}}|}\times 100\% \tag{19}$$

$$fld\text{-}F1=\frac{2|\{f|f\in F_{\text{true}}\bigcap F_{\text{infer}}\}|}{|F_{\text{true}}|+|F_{\text{infer}}|}\times 100\% \tag{20}$$

4.2 参数设置

如前所述，实验开始前，需要初始化推断约束的概率和字段边界判决阈值 *threshold*。在实践中，现有的概率推理文献通常使用先验知识来预设先验概率值，但由于概率推断算法往往进行了多轮迭代，因此推断的结果一般对这些先验概率值不敏感[4]。在对观察约束概率赋值时，分别取 0.9 和 0.1 表示可能和不可能。第一类推断约束概率应比第二类大，因为在某字节偏移确定为字段边界的情况下，该位置的特征大概率会满足针对字段边界设计的启发式规则，但反过来，有些满足这些规则的位置却不一定会是字段边界。结合实际情况，本文设置字节相似性第一类推断约束概率为 0.7，其余第一类设为 0.9；设置字节相似性第二类推断约束概率为 0.6，其余第二类设为 0.7。

threshold 的值通过每个偏移处 5 种观察约束概率的差异来确定。具体而言，首先判断每个位置观察约束概率大于或等于 0.9 的个数 *num*，筛选出所有 *num* 为 2 或 3 的位置，取这些位置的概率的中位数作为 *threshold*；如果没有 *num* 为 2 或 3 的位置，则取概率最小的 *num* 大于 3 的位置的概率作为 *threshold*；其余情况取 0.1。本次实验环境阈值取为 0.02，也就是说，概率大于 0.02 的偏移被初步判断为字段边界。

4.3 实验结果

选用常见的 3 种字段划分方法和本文方法进行比较，分别是 NMESYS[19]，ProSeg[20] 和 NETPLIER[4]，对比结果如表 1 所列。可以看出，除了在 s7comm_1k 和 modbus_1k 上，ProSeg 的准确率取值高于本文方案（加粗），NETPLIER 的召回率与本文方案持平（斜体），本文方案在其他所有指标上均表现最好。本文方案在两种规模的报文集上性能接近，所有协议都能取得 64%以上的准确率和 85%以上的召回率，其中 dnp3 协议的准确率超过 83%，召回率达到 95%。在边界 F1 值方面，本文方案在所有协议都超过了 75%，而要求最严苛的字段 $F1$ 值则超过了 38%。相对而言，ProSeg 的准确率总体表现优于召回率，出现了少划分的情况，边界 $F1$ 值在 70%左右，字段 $F1$ 值分布在 17%～43%之间；NETPLIER 则相反，出现了误划分，导致召回率的表现优于准确率，边界 $F1$ 值大部分在 73%以上，字段 $F1$ 值分布在 14%～50%之间；NMESYS 整体表现最差，准确率、召回率和边界 $F1$ 值均在 55%左右，大部分协议字段 $F1$ 值低于 20%，EtherNet/IP 协议取值甚至低于 10%。

表 1 字段划分方法对比

Table 1 Comparison of field segmentation methods

Protocol	Our Solution				NEMSYS				ProSeg				NETPLIER			
	P	*R*	*F*1	*fld*-*F*1	*P*	*R*	*F*1	*fld*-*F*1	*P*	*R*	*F*1	*fld*-*F*1	*P*	*R*	*F*1	*fld*-*F*1
s7comm_100	70.6	85.7	77.4	38.7	61.5	57.1	59.3	22.2	66.7	71.4	69.0	20.7	68.8	78.6	73.3	40.0
modbus_100	87.5	93.8	90.4	71.2	64.6	59.8	62.0	12.8	83.3	67.0	74.2	37.4	72.5	86.6	78.9	43.3
dnp3_100	83.3	95.0	88.7	75.3	58.5	52.3	55.1	14.6	73.9	66.8	70.0	35.0	73.9	80.9	77.2	49.9
EtherNet/IP_100	64.3	90.0	75.0	41.7	42.9	60.0	50.0	8.3	53.8	70.0	60.9	17.4	41.2	70.0	51.9	14.8
s7comm_1k	**75.0**	*85.7*	80.0	46.7	66.7	57.1	61.5	23.1	**78.6**	78.6	78.6	42.9	70.6	*85.7*	77.4	45.2
modbus_1k	**81.7**	*86.6*	84.0	52.1	62.5	67.0	64.6	12.8	**82.9**	66.1	73.3	28.3	72.5	*86.6*	78.9	43.3
dnp3_1k	83.3	95.0	88.7	75.3	58.5	52.3	55.1	14.6	81.3	61.8	70.2	32.5	73.9	80.9	77.2	49.9
EtherNet/IP_1k	64.3	90.0	75.0	41.7	42.9	60.0	50.0	8.3	53.8	70.0	60.9	17.4	41.2	70.0	51.9	14.8

通过对全部协议取平均结果，可得到 4 种评价指标的总体性能。图 5 给出了 4 种方法的总体性能，本文方案领先所有指标，在字段 F1 值的对比中优势尤其明显。3 种传统方法的对比中，NETPLIER 表现最佳，ProSeg 次之，NMESYS 最差。

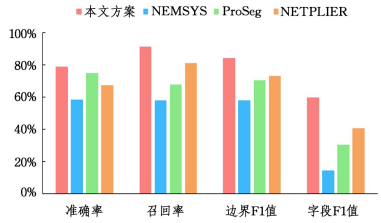

图 5 总体性能对比

Fig. 5 Overall performance comparison

除了同前人的工作进行对比，本文还比较了在 100 条报文上基于因子图的概率推断与投票专家算法的优劣。投票专家算法将 5 种约束特征看作 5 个投票者，每种特征根据约束概率决定是否投票，最后累计所有票数，判断票数是否超过阈值，超过阈值则推断该位置为边界，否则不是。因为 ProSeg 同样使用投票专家的方法，但其仅使用了涉及信息熵的 3 个特征，因此我们把 ProSeg 的结果也列入比较行列。如表 2 所列，本文方案在所有指标上都超过了投票专家算法，而投票专家算法比 ProSeg 多使用了两种特征，除了在 Ethernet/IP_100 协议上和 ProSeg 表现一致(加粗)，其他指标都超过了 ProSeg。实验结果表明更多维度的特征促进了边界识别，基于因子图的概率推断能综合利用所有特征，进一步提升性能。

表 2 边界判决方法对比

Table 2 Comparison of boundary inference methods

Protocol	Factor Graph				Voting Experts				ProSeg			
	P	*R*	*F*1	*fld*-*F*1	*P*	*R*	*F*1	*fld*-*F*1	*P*	*R*	*F*1	*fld*-*F*1
s7comm_100	70.6	85.7	77.4	38.7	68.8	78.6	73.3	33.3	66.7	71.4	69.0	20.7
modbus_100	87.5	93.8	90.4	71.2	85.7	80.4	82.9	55.7	83.3	67.0	74.2	37.4
dnp3_100	83.3	95.0	88.7	75.3	78.9	71.4	74.9	44.9	73.9	66.8	70.0	35.0
EtherNet/IP_100	64.3	90.0	75.0	41.7	**53.8**	**70.0**	**60.9**	**17.4**	**53.8**	**70.0**	**60.9**	**17.4**

本文方案使用同格式报文集作为输入，这种输入在实验环境下容易获取，但在现实环境中，通过聚类得到的报文集无法保证纯度。因此，我们设计了在 100 条 Ethernet/IP 报文中分别混入 1 条、5 条、10 条、15 条、20 条和 25 条 DHCP 报文的 6 种混淆报文集，用于测试本文方案的鲁棒性，实验结果如图 6、图 7 所示。

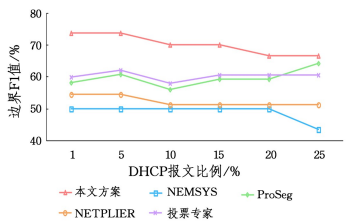

图 6 边界 F1 值鲁棒性对比

Fig. 6 Robustness comparison of boundary F1 score

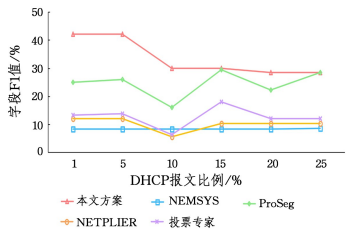

图 7 字段 F1 值鲁棒性对比

Fig. 7 Robustness comparison of field F1 score

通过实验可以发现，本文方案的边界识别和字段划分效果随着 DHCP 报文比例的增加而降低，但仍优于其他 4 种方法。ProSeg、NETPLIER 和投票专家 3 种方法表现有波动，其中 ProSeg 在混淆超过 15%后性能接近本文方案。NMESYS 由于是针对每条报文提取特征，虽然在 Ethernet/IP 协议上性能最差，但波动较小。

4.4 讨论分析

虽然本文方案在对比实验中表现较好，但仍存在一些不足。所有方法在 Ethernet/IP 协议上表现都较差，本小节以 Ethernet/IP 协议为例对实验结果进行讨论分析。本文使用的 Ethernet/IP 协议仅包含显性通信类型，它们的报文结构都一样，如图 8 所示。

封装头部	Command	Length	Session Handle	Status	Senden Context	Options
	2 字节	2 字节	4 字节	4 字节	6 字节	6 字节
封装数据	Interface Hande	Timeout	Item Count	Data		
	4 字节	2 字节	2 字节	若干字节		

图 8 Ethernet/IP 报文格式

Fig. 8 Message format of Ethernet/IP

根据 Ethernet/IP 报文结构，$O_{\text{true}}=\{2,4,8,12,20,24,28,30,32,-1\}$ 为真实边界。本文方案在 100 条报文集上

推断的字段边界为 $O_{infer}=\{1,3,4,8,12,14,15,16,20,24,28,30,32,-1\}$，其中－1 表示最后一个字节。图 9 给出了 8 条 Ethernet/IP 报文的前 32 字节，通过比较可发现，2 字节偏移处未被识别，而 1 和 3 字节偏移处却被推断为边界，原因是 Ethernet/IP 的报文采用的是小端模式，而 Command 和 Length 字段往往用不全 2 字节，后面的字节经常置为 0，因此在边界判定时容易把 Command 和 Length 字段首字节偏移处识别为字段边界。在 14，15，16 字节偏移处出现多余划分，这些字节都位于 Sender Context 字段，这个字段是由命令的发送者生成的随机 6 个字节值，因此容易导致误划分。此外，对齐填充符的使用也会导致误划分。要解决上述问题，需要研究如何增加数据分布多样性和降低大段随机字节与对齐填充符带来的影响。

70	00	3a	00	09	a6	18	3f	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	01	00	02	00
70	00	37	00	09	a6	18	3f	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	02	00
6f	00	22	00	09	a6	18	3f	00	00	00	00	1e	33	06	00	38	bc	27	0c	00	00	00	00	00	00	00	00	0a	00	02	00
6f	00	38	00	09	a6	18	3f	00	00	00	00	1e	33	06	00	38	bc	27	0c	00	00	00	00	00	00	00	00	04	02	00
70	00	3a	00	09	a6	18	3f	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	01	00	02	00
70	00	37	00	09	a6	18	3f	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	02	00
6f	00	22	00	09	a6	18	3f	00	00	00	00	25	33	06	00	38	bc	27	0c	00	00	00	00	00	00	00	00	0a	00	02	00
6f	00	38	00	09	a6	18	3f	00	00	00	00	25	33	06	00	38	bc	27	0c	00	00	00	00	00	00	00	00	04	02	00
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32

图 9 Ethernet/IP 报文前 32 字节

Fig. 9 First 32 bytes of messages of Ethernet/IP

结束语 本文提出了一种基于概率模型的二进制协议字段划分方案，首先从报文内在结构、报文间取值变化等 5 个维度构造字段边界约束关系，充分提取报文信息；其次在进行字段边界判决时，为避免简单的判决准则带来的偏差，以概率的方式将各种约束组合在一起，利用因子图模型计算各个位置成为边界的概率；最后基于边界概率进行边界判别并结合启发式规则进行修正。实验结果表明，本文方法在 4 个评价指标上的表现均优于 3 种传统的相关方法，同时具有较好的鲁棒性，能够对带有混淆报文的二进制协议进行字段划分。

本文方案仍然存在一些不足。首先，字段划分结果受限于样本空间的多样性；其次，大段随机字节和报文对齐引入的填充符会干扰字段边界的判别。这两个问题需要在未来的工作中进行进一步研究。

参 考 文 献

[1] SOPHOS. The Dirty Secrets of Network Firewalls [EB/OL]. (2018-04-10) [2021-06-10]. https://secure2. sophos. com/en-us/medialibrary/Gated-Assets/white-papers/firewall-dirty-secrets-report. pdf.

[2] WU L F, HONG Z, PAN P. Network Protocol Reverse Analysis and Application[M]. Beijing: National Defense Industry Press, 2016: 11-12.

[3] YAN X Y. Research and Implementation on the Key Technologies for Binary Private Protocol Reverse [D]. Zhengzhou: Strategic Support Force Information Engineering University, 2018.

[4] YE Y P, ZHANG Z, WANG F, et al. NETPLIER: Probabilistic Network Protocol Reverse Engineering from Message Traces [C]// 28th Annual Network and Distributed System Security Symposium. 2021.

[5] LIVSHITS B, NORI A V, RAJAMANI S K, et al. Merlin: Specification inference for explicit information flow problems [J]. ACM Sigplan Notices, 2009, 44(6): 75-86.

[6] COZZIE A, STRATTON F, XUE H, et al. Digging for Data Structures[C]// OSDI. 2008: 255-266.

[7] BEDDOE M A. Network protocol analysis using bioinformatics algorithms [J]. Toorcon, 2004, 26(6): 1095-1098.

[8] CUI W, PAXSON V, WEAVER N, et al. Protocol-Independent Adaptive Replay of Application Dialog[C]// 13th Annual Network and Distributed System Security Symposium. 2006.

[9] CUI W, KANNAN J, WANG H J. Discoverer: Automatic Protocol Reverse Engineering from Network Traces[C]// USENIX Security Symposium. 2007: 1-14.

[10] KRUEGER T, KRÄMER N, RIECK K. ASAP: Automatic semantics-aware analysis of network payloads[C]// International Workshop on Privacy and Security Issues in Data Mining and Machine Learning. Springer, 2010: 50-63.

[11] PAN F, HONG Z, DU Y X, et al. Recursive Clustering Based Method for Message Structure Extraction [J]. Journal of Sichuan University (Engineering Science Edition), 2012, 44(6): 137-142.

[12] WANG Y, LI X, MENG J, et al. Biprominer: Automatic mining of binary protocol features[C]// 2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies. IEEE, 2011: 179-184.

[13] WANG Y, YUN X, SHAFIQ M Z, et al. A semantics aware approach to automated reverse engineering unknown protocols [C]// 2012 20th IEEE International Conference on Network Protocols (ICNP). IEEE, 2012: 1-10.

[14] LI M, YU S Z. Noise-Tolerant and optimal segm-entation of message formats for unknown application-layer protocols [J]. Journal of Software, 2013, 24(3): 604-617.

[15] 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. ACM, 2014: 51-62.

[16] ZHANG Z, ZHANG Z, LEE P P C, et al. ProWord: An unsupervised approach to protocol feature word extraction[C]// IEEE INFOCOM 2014－IEEE Conference on Computer Communications. IEEE, 2014: 1393-1401.

[17] BERMUDEZ I, TONGAONKAR A, ILIOFOTOU M, et al. Automatic protocol field inference for deeper protocol understanding[C]// 2015 IFIP Networking Conference (IFIP Networking). IEEE, 2015: 1-9.

[18] BERMUDEZ I, TONGAONKAR A, ILIOFOTOU M, et al. Towards automatic protocol field inference [J]. Computer Communications, 2016, 84: 40-51.

[19] KLEBER S, KOPP H, KARGL F. {NEMESYS}: Network Message Syntax Reverse Engineering by Analysis of the Intrinsic Structure of Individual Messages[C]// 12th {USENIX} Workshop on Offensive Technologies({WOOT} 18). 2018.

[20] SUN F H, WANG S, ZHANG C R, et al. Unsupervised field segmentation of unknown protocol messages [J]. Computer Communications, 2019, 146: 121-130.

[21] JIANG D, LI C, MA L, et al. ABInfer: A Novel Field Boundaries Inference Approach for Protocol Reverse Engineering[C]// 2020 IEEE 6th International Conference on Big Data Security on Cloud(Big Data Security), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security(IDS). IEEE, 2020: 19-23.

[22] WANG X, LV K, LI B. IPART: an automatic protocol reverse engineering tool based on global voting expert for industrial protocols[J]. International Journal of Parallel, Emergent and Distributed Systems, 2020, 35(3): 376-395.

[23] LIU O, ZHENG B, SUN W, et al. A Data-driven Approach for Reverse Engineering Electric Power Protocols [J]. Journal of Signal Processing Systems, 2021, 93(Jan): 1-9.

[24] KATOH K, MISAWA K, KUMA K, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform [J]. Nucleic Acids Research, 2002, 30(14): 3059-3066.

[25] KATOH K, STANDLEY D M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability [J]. Molecular Biology and Evolution, 2013, 30(4): 772-780.

[26] KLEBER S, MAILE L, KARGL F. Survey of protocol reverse engineering algorithms: Decomposition of tools for static traffic analysis [J]. IEEE Communications Surveys & Tutorials, 2018, 21(1): 526-561.

[27] SHLEZINGER N, FARSAD N, ELDAR Y C, et al. Data-driven factor graphs for deep symbol detection[C]// 2020 IEEE International Symposium on Information Theory(ISIT). IEEE, 2020: 2682-2687.

[28] GIENGER A, SAWODNY O. Data-based Process Monitoring and Iterative Fault Diagnosis using Factor Graphs[C]// 2020 IEEE International Conference on Industrial Technology (ICIT). IEEE, 2020: 35-40.

[29] KOTIANG S, ESLAMI A. Boolean factor graph model for biological systems: the yeast cell-cycle network[J]. BMC bioinformatics, 2021, 22(1): 1-27.

[30] LEANZA A, REINA G, BLANCO-CLARACO J L. A Factor-Graph-Based Approach to Vehicle Sideslip Angle Estimation [J]. Sensors, 2021, 21(16): 5409.

[31] ANKAN A, PANDA A. pgmpy: Probabilistic graphical models using python[C]// Proceedings of the 14th Python in Science Conference(SCIPY). 2015: 6-11.

YANG Zi-ji, born in 1994, postgraduate. His main research interests include protocol reverse engineering and network traffic classification.

ZHU Yue-fei, born in 1962, Ph. D, professor, Ph. D supervisor. His main research interests include information security and public key cryptography.

(责任编辑:何杨)