计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190700031

基于最长连续间隔的未知二进制协议格式推断

陈庆超[1]　王　韬[1]　冯文博[2]　尹世庄[1]　刘丽君[1]
1 陆军工程大学装备模拟训练中心　石家庄 050003
2 陆军工程大学指挥控制工程学院　南京 210007
(cqc62808@163.com)

摘　要　在未知二进制协议的格式推断过程中,常常引入大量的先验知识,实验操作复杂且准确率不高。为此,文中提出了一种人为设定较少参数、操作简单、准确率较高的方法进行未知二进制协议格式推断,将预处理的协议数据进行层次聚类,以CH(Calinski-Harabasz)系数为评价标准获得最优聚类,通过对聚类所得结果进行改进的序列对比以获得带有间隔的协议数据序列,统计合并连续间隔,以分析协议格式。实验结果表明,提出的二进制协议格式推断方法能够推断出未知二进制协议80%以上的字段间隔,相较于AutoReEngine算法中的格式推断方法,所提方法的F1-Measure值整体上提升了约30%。

关键词:二进制协议;格式推断;层次聚类;序列对比;间隔
中图法分类号　TP393

Unknown Binary Protocol Format Inference Method Based on Longest Continuous Interval

CHEN Qing-chao[1],WANG Tao[1],FENG Wen-bo[2],YIN Shi-zhuang[1] and LIU Li-jun[1]
1 Equipment Simulation Training Center,Army Engineering University,Shijiazhuang 050003,China
2 College of Command and Control Engineering,Army Engineering University,Nanjing 210007,China

Abstract　In the process of format inference of unknown binary protocols,a large amount of prior knowledge is often introduced,the experimental operation is complex and the accuracy of the results is low. For this reason,a method that requires less artificial setting of parameters,simple operation and higher accuracy is proposed to infer the unknown binary protocol format. The preprocessed protocol data is clustered hierarchically,and the optimal clustering is obtained by using CH (Calinski-Harabasz) coefficient as the evaluation criteria. Through the improved sequence comparison of the clustering results,the protocol data sequence with interval is obtained,continuous intervals are counted and merged to analyze protocol formats. The experimental results show that the binary protocol format inference method proposed in this paper can infer more than 80% of the field intervals in the unknown binary protocol. Compared with the format inference method in AutoReEngine algorithm,the F1-Measure value of the proposed method is improved by about 30% as a whole.

Keywords　Binary protocol,Format inference,Hierarchical clustering,Sequence alignment,Interval

1　引言

二进制协议因其传输高效且冗余较少,被广泛应用于物联网和僵尸网络的控制中。出于安全、利益等考虑,细节不公开的二进制协议被广泛使用,这些不公开的协议给网络的安全和控制带来了很大的挑战。对未知二进制协议进行逆向分析是现有相关研究的重要课题。

传统的针对未知协议的逆向分析主要分为基于网络流量的分析和基于指令执行轨迹的分析。基于指令执行轨迹的分析方法虽然具有较高的准确率,且能获得协议的语义分析,但依赖于协议实现终端,通用性不强。而基于网络流量的分析方法以网络流量为研究对象,相对比较灵活,但准确率较低,且对协议的语义不敏感[1]。本文采用基于网络流量的分析方法重点研究二进制协议的格式推断。

目前,已有一些针对二进制协议格式推断的研究成果。Luo等提出AutoReEngine方法进行协议格式推断,该方法在提取频繁序列之后,根据位置信息筛选频繁序列,并进行格式以及状态机的推断[2]。Li等提出基于隐半马尔可夫模型(Hidden Semi-Markov Models,HSMM)的协议格式分段方法,该方法利用报文序列样本估计模型参数,并利用基于HSMM的最大似然概率分段方法对报文各字段进行划分,能对噪声进行有效过滤,但其依赖于会话流[3]。Zhang等提出

到稿日期:2019-07-03　返修日期:2019-09-21　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划(2017YFB0802900);江苏省自然科学基金(BK20161469)
This work was supported by the National Key Research and Development Program of China (2017YFB0802900) and Natural Science Foundation of Jiangsu Province, China (BK20161469).
通信作者:王韬(a13592247640@foxmail.com)

利用基于信息熵的专家投票系统进行候选协议关键词提取的ProWord方法，该方法在获得候选关键词之后，使用排名算法筛选高排名的候选关键词作为协议的关键词[4]。Tao等提出PRE-Bin方法进行以bit为最小单位的格式推断，该方法引入轮廓系数进行聚类，使用改进的NW(Needleman-Wunsch)[5]算法进行序列比对，然后利用贝叶斯推断判定字段边界，但精度不高[6]。Yan等提出基于最佳路径搜索的二进制协议格式关键词边界确定方法，该方法将位置因素引入n-gram算法，有效解决了n-gram算法中n值的确定问题[7]。Wang等提出基于关键词递归循环提取的Bioprominer方法，并建立了状态转移模型[8]。Hou等提出基于位置的自动化网络流协议逆向方法PoKE，从关键词的角度出发，同时为关键词添加位置属性，从而可以提取出较短的关键词[9]。Liu等提出基于有损计数算法改进的投票专家算法来划分协议关键字[10]。Meng等提出基于层级聚类和概率对齐的协议逆向方法[11]。总结现有研究发现，多数研究是基于频繁序列挖掘提取关键词以推断协议格式，虽然文本协议的关键词不变，但二进制协议的固定位置没有固定的关键词，即意味着通过频繁序列挖掘提取二进制协议的关键词需要设定更低的频繁度阈值，而低频繁度阈值会引起关键词序列和用户数据难以区分的问题。

针对现有研究存在的不足，本文提出一种不依赖会话流的二进制协议格式推断算法，能够高效地划分二进制协议格式。该算法首先使用改进的层次聚类算法对协议数据进行有效聚类，针对二进制协议格式对齐的特点，使用改进的序列对比算法，对聚类后的协议数据进行序列对比，在此基础上统计合并最长连续间隔，以获取协议格式信息。

本文的主要工作如下：1)提出了改进的层次聚类方法对二进制协议报文进行聚类，使得聚类过程清晰，易于人为分析，聚类方法使用汉明距离衡量报文之间的距离；使用CH系数评估聚类效果，能有效地对二进制协议进行相似性聚类。2)提出了改进的序列对比方法，在序列对比的过程中不会导致二进制协议的格式错位，能有效地找出序列之间的不同之处。3)提出了利用统计序列对比结果中的最长间隔来推断未知二进制协议格式间隔的方法，能够高效地推断出未知二进制协议的格式。

2 基于最长连续间隔的未知二进制协议格式推断

无论是文本协议还是二进制协议，都具有一定的协议格式，即协议的不同字段代表不同的协议状态，具有不同的属性。本文研究的主要内容是，根据统计分析对捕获的二进制协议数据进行格式上的推断，以便对协议状态、控制等进行进一步的分析。本文重点解决三大问题：1)二进制协议数据的相似性聚类；2)二进制协议数据的序列对比；3)基于以上两点的协议格式推测。

2.1 研究对象

本文的研究对象是作为已知协议载荷的二进制协议数据，如BITTORRENT协议。相较于文本协议，如HTTP协议，二进制协议的格式部分往往具有固定位置和固定长度的特点，但不具有类似换行符等的分隔符，且字段往往不是使用文本而是使用数字进行描述，因此协议的格式并不是一目了然的，需要对其进行格式上的推断。在获取一定数量具有相同格式的协议数据的前提下，根据统计学规律可使协议格式推断成为可能。

2.2 问题定义

基于网络流量的二进制协议格式推断主要使用统计学方法，即根据局部性原理，相似的时间和空间产生的连续的网络流往往具有相似的格式、相近的语义。但是对于不公开的未知二进制协议，在没有任何先验信息的前提下，由于各字段的长度以及字段值具有不确定性，在进行格式推断时，很容易造成格式边界模糊，增加格式推断的难度。

2.3 方法概述

本文提出的二进制协议格式推断方法主要分为4个阶段：预处理阶段、层次聚类阶段、序列对比阶段以及最长连续间隔统计阶段，如图1所示。其中，预处理阶段的主要工作是对原始数据报文进行收集、切分以及去重，以获得实验数据集合。层次聚类阶段根据相似性对实验集合进行聚类，以获得属于不同类别的小集合。序列对比阶段对各个小集合分别进行序列对比。最长连续间隔统计阶段对序列对比中获得的所有比对结果进行统计，进而对协议格式进行推断。

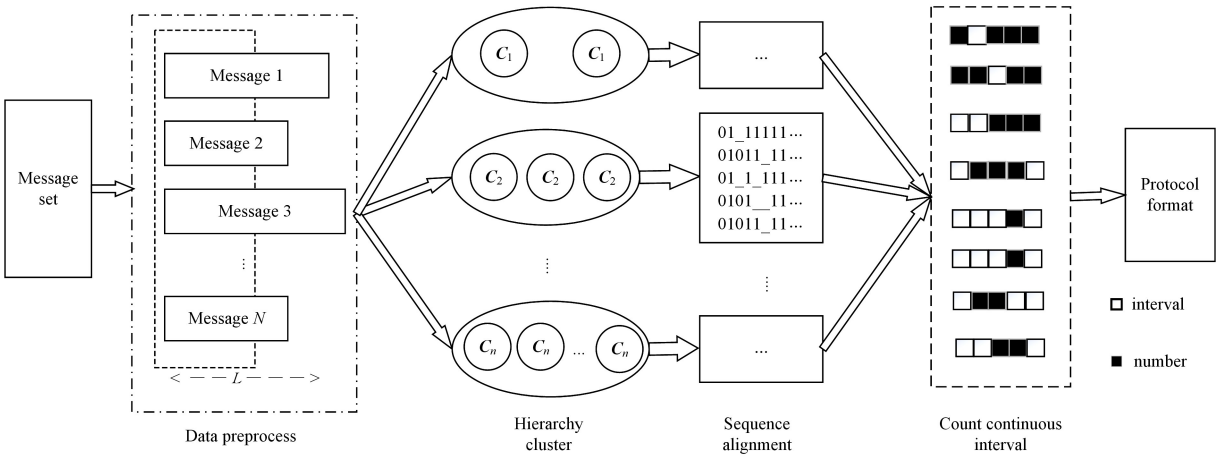


图1 二进制协议格式推断的流程图

Fig. 1 Flow chart of binary protocol format inference

3 预处理阶段

二进制协议数据主要分为两个部分，即协议头和协议载荷。协议头部相对较短，而协议载荷长度不固定，参考 Li 等提出的基于报文长度截取的数据预处理算法[12]，初步剔除协议数据中的大部分载荷片段，具体做法为：设置长度阈值 L，保留协议数据头部 L 个字节的数据，代替原有的报文，然后去除实验数据中的重复数据，以提高协议格式推断的效率。

长度阈值 L 的确定可以参考汉明重量的方差设定[13]。在二进制协议的协议字段中，相同位置的字段分块的汉明重量统计信息相对固定，因此汉明重量的方差也相对较小。而相对于协议字段而言，具有不确定取值的用户数据使得固定分块长度内的用户数据的汉明重量也同样更具有不确定性，进而汉明重量的方差也相对更大，因此可以根据汉明重量的方差划分二进制协议中的格式字段和用户数据字段。

4 层次聚类阶段

针对二进制协议报文的特点，本文提出一种使用汉明距离作为距离度量，使用 CH 系数判定聚类簇数的层次聚类算法，对经过预处理的报文数据进行聚类，算法的步骤为：

1）获取所有样本的距离矩阵；

2）将每个数据点作为一个单独的簇；

3）合并两个最接近的簇；

4）更新样本的距离矩阵；

5）重复步骤 2）－4），直到所有的样本都属于同一个簇；

6）根据树状图，设置切分距离的变化域；

7）在变化域内，根据距离对聚类结果进行切分，计算 CH 系数；

8）取获得最大 CH 系数的距离值作为最优距离，以获得最佳聚类。

报文之间的距离采用汉明距离来衡量，如式(1)所示：

$$d(U,V)=\frac{\sum_{i=0}^{n}(U[i]\neq V[i])}{n} \tag{1}$$

其中，U，V 分别代表两个报文序列，n 为报文序列长度。

本文的聚类算法引入了 CH 系数来评估聚类效果，如式(2)所示：

$$CH(k)=\frac{\mathrm{tr}(\boldsymbol{B}_k)m-k}{\mathrm{tr}(\boldsymbol{W}_k)k-1} \tag{2}$$

其中，m 为训练样本数，k 为类别数，$\boldsymbol{B}_k$ 为类别间的协方差矩阵，$\boldsymbol{W}_k$ 为类别内部数据的协方差矩阵，tr 为矩阵的迹。使用汉明距离衡量报文之间的距离，降低了不同类型的报文划分到同一簇的概率。

算法伪代码如算法 1 所示。

算法 1 改进的层次聚类算法

输入：Dataframe of protocol data in 4 bit amax

输出：CH 系数集合 chs1

```
1. Z= hierarchy. linkage(amax, method='average', metric='hamming') /* 使用 hamming 距离作为度量 */
2. for j in range(low, high) do /* 设置变化域范围 */
3. clusters=fcluster(Z,j,criterion='distance')/* 调用层次聚类库函数 */
4. chs1. append(metrics. calinski_harabaz_score(amax, clusters)) /* 计算 CH 系数 */
5. end do
6. return chs1
```

相比于其他类型的聚类算法，如 K-means 聚类算法，本文的聚类算法基于层次聚类，继承了层次聚类的相关优势，不需要预先设置参数，减少了人工参与，降低了人为因素带来的误差，同时，使用汉明距离衡量报文之间的距离，增加了聚类的精确度；使用 CH 系数评估聚类效果，使得划分聚类簇数更具有可信性。

经过层次聚类，将各个数据报文根据相似性聚类成簇，为下一步将相似的数据报文进行序列对比以获取间隔段提供了支持。

5 序列对比阶段

序列对比也称为模式匹配，根据一次匹配的模式个数，可分为单模式匹配和多模式匹配[14]。单模式匹配一次查找单个模式出现的位置，多模式匹配一次查找多个模式出现的位置。

5.1 模式匹配算法

常用的单模式匹配算法主要有 BF(Brute Force)算法、KMP(Knuth-Morris-Pratt)算法、BM(Boyer-Moore)算法等。单模式匹配算法一次只能查找一个模式串，因此效率不高。多模式匹配常采用复杂的数据结构对数据状态进行存储，常用的多模式匹配算法主要有 AC(Aho-Corasick)算法[15-16]、AC-BM 算法和 WM(Wu-Manber)算法等。多模式匹配算法一次可以匹配多个模式，效率较高，但实现稍微复杂。未知协议的频繁模式挖掘常采用多模式匹配算法，如 Lei 等采用基于 AC 算法进行未知帧的前导码挖掘[17]，Li 等在评分函数上对 SW(Smith-Waterman)[18]算法和 NW 算法进行改进，以获得更精确的序列对齐[19]。

5.2 改进的序列对比算法

多序列对比算法最初应用于生物科学中遗传信息格式的比对，而应用于二进制协议格式的比对时，由于二进制协议数据的位置对齐的特点，如果使用传统的序列对比方法插入间隔，可能会造成二进制协议数据虽然获得了最大相似性匹配，但失去了位置对齐的优势。因此对于二进制协议，序列对比应尽量减少添加间隔造成的影响。文献[6]在 NW 算法的评分函数和替换机制上进行改进，即一方面增加间隔的罚分至－2，另一方面将间隔的插入改为使用间隔替换原有符号，以获得更合理的目标序列，经过改进后的序列对比没有破坏二进制协议相较于文本协议独有的结构属性。虽然该方法试图降低间隔插入的影响，但是并没有解决传统的序列对比算法罚分的人为设定对序列对比结果的影响，罚分的设定是传统序列对比算法中引入的对实验结果有很大影响的先验知识。

为了降低间隔产生的影响以及避免对罚分策略的合理性进行解释，受文献[6]采用改进的序列对比算法的启发，本文提出进一步改进的序列对比算法。所提算法中，经过层次聚类之后获得的各个簇中的数据是相似的，指定一个序列作为模板序列，其他作为待比对序列，以 4 bits 为最小比对单元，

将待比对序列和模板序列进行一一对比，使用间隔替换与模板序列中不匹配的数据，得到一个存在间隔的数据序列，如图 2 所示。

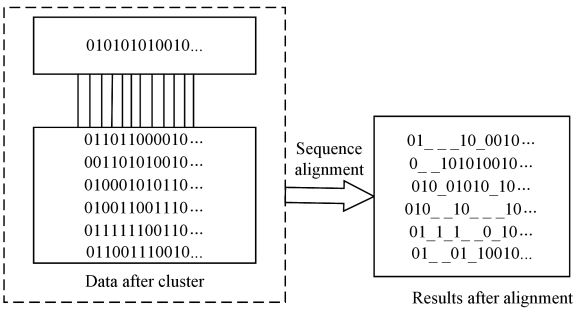


图 2 改进的序列对比

Fig. 2 Improved sequence alignment

具体执行步骤为如下。

1) 随机在聚类后的一类数据中选出一条数据作为模板序列 $sequence1=\langle b_1 b_2 b_3 \cdots b_n \rangle$，其中 b_i 为在 i 位置上的数值，n 为序列长度。

2)将属于同一簇中的一条协议数据 $sequ=\langle a_1 a_2 a_3 \cdots a_n \rangle$ 作为待比对序列，其中 a_i 为在 i 位置上的数值，将其与模板序列相同位置上的数值进行一一对比，如果待比对序列中与模板序列在相同位置上的数值不同，则将该位置数值替换为间隔，例如若 a_i 不等于 b_i，则将 a_i 的值替换为间隔“_”。

3) 重复步骤 2)，直至属于同一簇的数据与模板序列比对完全。

改进的序列对比算法的伪代码如算法 2 所示。

算法 2 改进的序列对比算法

```
输入:模板序列 sequence1,其他序列 sequence2
输出:带有间隔的序列 sequence3
1. length=len(sequence1)
2. for j in range(length) do
3.    if (sequence1[j]==sequence2[j])
4.    sequence3=sequence3+sequence1[j]
5.    else
6.    sequence3=sequence3+'_'
7. end do
8. return sequence3
```

序列对比方法，如 NW 算法等，最早被应用于 DNA 等遗传信息的分析，后被应用于文本协议的格式推断，而算法的使用目的是使文本协议根据关键词进行格式上的对齐，即以关键词划分格式。二进制协议不存在类似 GET 的关键词，且本身格式对齐，因此以格式对齐为目的对未知二进制协议进行序列对比是不恰当的。

本文提出的序列对比算法消除了人为设定罚分对序列对比结果造成的影响，且得到的结果保持了二进制协议的结构特点，消除了人为因素带来的误差对进一步的格式推断的影响。

6 最长连续间隔统计阶段

常见的划分格式方式主要分为两类，一类依靠报文各字段内部的统计特性来划分，如 VDV(Variance of the Distribution of Variances)[20]、频繁序列；另一类依靠报文各字段间的差异划分，如信息熵[21]。基于统计的频繁序列的主观人为因素在于支持度的设置，此类方法容易筛除一部分出现频率低于最小支持度的项，而 VDV 和信息熵则忽视了字段内部本身存在的差异，如数字的百位和个位本身的独立以及其统计特性上的差异。但是对于一个非随机数字逐渐增大特性的字段，一般低位的变化率相比高位的变化率较大，这种特性一方面导致了字段内部的统计特性不一致，另一方面导致了字段低位数据统计特征更加显著，即意味着如果将关注点从传统方法中的对整个字段的挖掘向挖掘字段的低位部分倾斜，由于最低位的后边界也是一个字段的后边界且上一个字段的后边界也是下一个字段的前边界，那么挖掘出一个包含低位数据的字段片段就能将两个字段分离。这种字段内部统计特性的不一致也导致了基于频繁序列挖掘二进制协议的低准确率。

本文根据二进制协议格式的特点，提出统计最长间隔段的方法对经过改进的序列对比处理后的数据进行进一步处理。经过改进的序列对比处理之后，可得到带有间隔段的数据序列，其组成间隔段集合，然后统计所有数据序列中间隔的位置。间隔段之间的关系主要包括 3 种：相离、相交、包含，如图 3 所示。

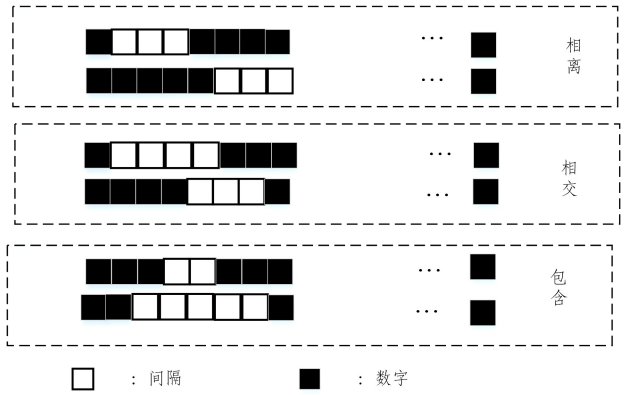


图 3 相离、相交、包含

Fig. 3 Separation, intersection, inclusion

对集合进行压缩，首先删除所有被包含的间隔段，具体思路如下：

1)设起始查找位置为 I；

2)向 I 前方查找至 P 位置，使集合中不存在 P 至 I 连续的间隔段，再向后查找至 Q 位置，使集合中不存在 I 到 Q 连续的间隔段，记录间隔 $Interval_{(I+1,Q-1)}$；

3)起始位置更改为 Q；

4)重复步骤 2)— 3)至序列末尾位置。

算法复杂度分析：将集合中的间隔所在位置集合映射为数字集合，对集合的压缩思路就转化为数字串查找，即从 1 开始在数字集合中查找连续数字串，时间复杂度为 $O(L \times N)$，其中 L 为长度阈值，N 为间隔段集合中间隔的总长度。

对于两个间隔段相交现象，更为合理的处理方法是将相交部分划分到前一个间隔段，如果将其划分到后一个间隔段，则第一个间隔段就包含了一个字段的尾部和另一个字段的头部，不符合字段内部从低位向高位变化的规律。该方法时间复杂度不变。

同时，为使算法更具稳定性，可以设定一定的阈值，将间隔段中的间隔出现较少的部分片段从该间隔段中删除，以避免因为数据不纯以及聚类算法效果不理想带来的误差。

本文提出的通过统计最长连续间隔获取字段边界的方法，最大化地减少了人为因素的干扰；同时考虑了协议内部各字段的变化规律，即虽然字段内部不同位置的变化规律不同，但是字段的变化频率一般是低位高于高位，且字段从低位向高位变化，这避免了字段内部因统计规律不一致带来的误差；对于协议数据的微小变化较敏感，本文方法更容易获取变化不频繁的字段的协议格式。

7 仿真与实验分析

7.1 实验方法

为了对本文所提方法进行验证，选取二进制协议 ARP（Address Resolution Protocol），DNS（Domain Name System），ICMP（Internet Control Message Protocol）和 S7COMM（S7 Communication）作为研究对象进行测试，实验数据来源于互联网开源数据[22-23]，测试数据集如表 1 所列。

表 1 测试数据集

Table 1 Testing data sets

Protocol type	Number of frame
ARP	1 000
DNS	5 666
ICMP	769
S7COMM	304

本文采用精确率（precision）、召回率（recall）和 F1-Measure 值作为衡量指标对实验结果进行评估。设本文实验中确定的字段边界集合为 $W=\{w_1,w_2,w_3,\cdots,w_p\}$，真实的字段边界为 $Y=\{y_1,y_2,y_3,\cdots,y_q\}$，精确率、召回率和 F1-Measure 的计算如式（3）－式（5）所示：

$$precision=\frac{card(W\bigcap Y)}{card(W)} \tag{3}$$

$$recall=\frac{card(W\bigcap Y)}{card(Y)} \tag{4}$$

$$F1\text{-}Measure=\frac{2\times precision\times recall}{precision+recall} \tag{5}$$

其中，$card(A)$ 为集合 A 中元素的数目。

7.2 实验与分析

本节对本文所提算法的整体性能进行了测试和对比，为了简化实验，只选取二进制协议格式中的二进制字段作为实验数据，并将实验数据中从未变化过的多个连续字段视为一个字段。针对数据集中 4 种不同类型的协议数据，分别计算其精确率、召回率和 F1-Measure 值，实验结果如图 4 所示。

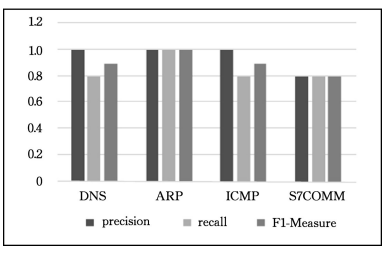


图 4 整体性能

Fig. 4 Overall performance

由实验结果可知，本文所提方法可以大致提取出测试协议所有的字段边界（精确率体现），并且提取的无关字段边界占比较少（召回率体现），其中部分字段如描述协议状态的字段未能正确提取出字段边界，其一方面与这些字段内部各个数据的独立性有关，即字段内部又可以划分成小字段，代表不同的属性；另一方面也与聚类算法的不精确相关，导致不能将类型相似的数据精确区分。实验数据中 S7COMM 协议属于工控协议，表明了本文所提算法在工控协议逆向分析中的有效性。

选择文献[2]提出的 AutoReEngine 算法作为对比算法，该算法基于关键词推断协议格式且以字节为基本单位。基于关键词推断协议格式是建立在频繁序列挖掘的基础之上，根据支持度等进行序列挖掘容易对序列之间的微小差异不敏感，且需要大量的数据集作为实验支撑。基于关键词推断协议格式无法避免的问题是关键词频繁度的设置，实验之前无法对数据进行整体上的分析。对于变化不频繁、相对单一的数据，需要设置较高的频繁度，而对于变化频繁、种类繁多的数据，需要设置较低的频繁度，易导致关键词之间的频繁度差异不明显。以上这些问题都有可能导致挖掘出的关键词之间存在连接的问题。对于文本协议，由于关键词数值固定，这种连接现象不明显，但是对于二进制协议，则很容易出现划分的结果并不是一个完整的关键词，而是一个关键词和另一个关键词首部的连接。对于变化频繁的数据，需要设置较低的频繁度，导致数据和关键词之间的频繁度不明显，不易区分。因此本文所提方法相较之下具有更高的精确度。在 4 种协议集上，对本文方法与 AutoReEngine 算法的 F1-Measure 值进行统计比较，结果如图 5 所示。

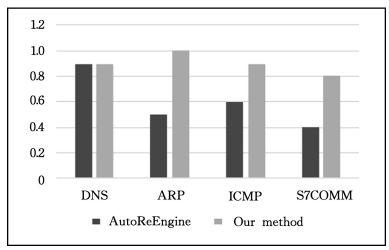


图 5 算法性能对比

Fig. 5 Comparison of algorithm performance

通过上述实验结果分析，本文所提方法能够推断出未知二进制协议中 80％以上的字段边界，且相较于 AutoReEngine 算法，本文所提方法的 *F1-Measure* 值整体上提升了约 30％。本文所使用的数据量较少，但是仍然可以保持较好的实验结果，而基于频繁序列提取关键词的 AutoReEngine 算法需要大量的数据支撑，在较少的数据集下效果明显下降，这是基于关键词推断未知协议格式的方法普遍存在的问题。

基于频繁序列的格式推断方法的着眼点在于部分数据的重复出现，而本文方法的着眼点在于固定格式下数据之间变化频率的差异，即部分数据较高的变化率。基于频繁序列的格式推断方法需要大量的额外存储空间以对所有不同长度的序列进行计数，而本文所提方法是对固定位置上的数据的变化进行统计，需要的存储空间较小。时间复杂度上，本文方法的序列对比阶段和最长连续间隔查找阶段的时间复杂度均为 O(N)，在运行时间上有一定的保证。

结束语 本文提出了一种基于统计最长连续间隔的二进制协议格式推断方法，通过改进的层次聚类对预处理的二进制协议数据进行聚类，在此基础上对聚类后的数据进行改进的序列对比，最后统计序列对比后的最长连续间隔，得出大致的字段边界。在 ARP，DNS，ICMP 和 S7COMM 4 种不同类型的二进制协议数据集上的测试结果表明，本文所提方法具有一定的推断二进制协议格式的能力。本文只是对最长连续间隔交叉现象进行简单的合并处理，并没有对算法的容错能力进行测试。根据统计规律对交叉现象进行进一步处理，以及对所得边界进行循环验证是未来的研究课题。

参 考 文 献

[1] DUCHENE J, LE GUERNIC C, ALATA E, et al. State of the art of network protocol reverse engineering tools[J]. Journal of Computer Virology and Hacking Techniques, 2018, 14(1): 53-68.

[2] LUO J Z, YU S Z. Position-based automatic reverse engineering of network protocols[J]. Journal of Network and Computer Applications, 2013, 36(3): 1070-1077.

[3] LI M, YU S Z. Noise-Tolerant and Optimal Segmentation of Message Formats for Unknown Application-Layer Protocols [J]. Journal of Software, 2013(3): 604-617.

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

[5] MUHAMAD F N, AHMAD R B, ASI S M, et al. Performance Analysis Of Needleman-Wunsch Algorithm (Global) And Smith-Waterman Algorithm (Local) In Reducing Search Space And Time For Dna Sequence Alignment[C]// Journal of Physics: Conference Series. IOP Publishing, 2018, 1019(1): 012085.

[6] TAO S, YU H, LI Q. Bit-oriented format extraction approach for automatic binary protocol reverse engineering[J]. IET Communications, 2016, 10(6): 709-716.

[7] YAN X, LI Q. Method for determining boundaries of binary protocol format keywords based on optimal path search[J]. Journal of Computer Applications, 2018, 38(6): 1726-1731.

[8] WANG Y, LI X, MENG J, et al. Biprominer: Automatic Mining of Binary Protocol Features[C]// International Conference on Parallel & Distributed Computing. IEEE, 2012: 179-184.

[9] HOU F J, WANG L, WANG S, et al. Position-based Automated Protocol Reverse Engineer on Network Flows[J]. Computer Engineering, 2019, 45(5): 84-87.

[10] LIU J L, FU G Y, LI H L, et al. Proprietary protocol fuzzing method based on improved voting expert algorithm[J]. Computer Engineering and Applications, 2018, 54(12): 98-104.

[11] MENG F, ZHANG C, WU G. Protocol reverse based on hierarchical clustering and probability alignment from network traces [C]// 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA). IEEE, 2018: 443-447.

[12] LI Y, LI Q, ZHANG X. Automatic protocol format signature construction algorithm based on discrete series protocol message [J]. Journal of Computer Applications, 2017, 37(4): 954-959.

[13] WU Y. Research on Encryption Identification and frequent patterns mining of unknown protocol bitstreams[D]. Shijiazhuang: Army Engineering University, 2015: 130-132.

[14] ASHKENAZY H, SELA I, LEVY KARIN E, et al. Multiple sequence alignment averaging improves phylogeny reconstruction [J]. Systematic Biology, 2018, 68(1): 117-130.

[15] HASHEEM Y M, MOHAMAD K M, ABDI A N E, et al. Mobile Forensic Images and Videos Signature Pattern Matching using M-Aho-Corasick [J]. International Journal of Advanced Computer Science and Applications, 2016, 7(7): 261-264.

[16] QIAO Z, GOTO K, OHSHIMA T, et al. Dictionary matching: review of the aho-corasick algorithm and vision for large dictionaries[C]// Proceedings of the 8th International Conference on Information Systems and Technologies. ACM, 2018: 4.

[17] LEI D, WANG T, WANG X H, et al. Unknown protocol frame segmentation algorithm based on preamble mining [J]. Journal of Computer Applications, 2017, 37(2): 440-444.

[18] LIAO Y L, LI Y C, CHEN N C, et al. Adaptively Banded Smith-Waterman Algorithm for Long Reads and Its Hardware Accelerator[C]// 2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP). IEEE, 2018: 1-9.

[19] LI T, LIU Y, ZHANG C, et al. A noise-tolerant system for protocol formats extraction from binary data [C] // 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA). IEEE, 2014: 862-865.

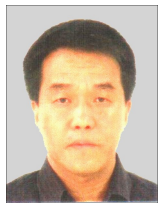
[20] TRIFILO A, BURSCHKA S, BIERSACK E. Traffic to protocol reverse engineering[C]// 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. IEEE, 2009: 1-8.

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

[22] WRCCDC: Pcaps from the Western Regional Collegiate Cyber Defense Competition [OL]. https://archive.wrccdc.org/pcaps/.

[23] CSDN. S7 协议数据集[OL]. https://download.csdn.net/download/jizhuan0248/10780517.

CHEN Qing-chao, born in 1996, postgraduate. His main research interests include cyber security and so on.



WANG Tao, born in 1964, Ph.D, professor. His main research interests include cyber security and cryptography.