•计算机科学与技术•

DOI:10.12454/j.jsuese.202301023

本刊网刊

基于稳态属性的工业控制协议操作字段识别

覃　朗[1,2],陈兴蜀[2,3*],朱　毅[2,3],李　尧[2,3],何　军[1]

(1.四川大学 计算机学院,四川 成都 610064;2.四川大学 网络空间安全研究院,四川 成都 610065;3.四川大学 网络空间安全学院,四川 成都 610065)

摘　要:工业控制协议中的操作字段是刻画识别工控网络行为、理解和监控网络活动的基础和关键数据,在工控网络流量中对操作字段进行识别抽取具有重要意义。然而,目前的操作字段识别大多为依赖专家经验知识的人工分析提取方法,存在效率不高、通用性不足且不能处理工控领域中大量存在的、未公开的私有协议的缺陷,无法在场景、协议等未知的复杂网络情况下自动识别出操作字段。为解决上述问题,本文利用工控网络独有的领域特征,跳出协议和程序的限制,提出一种基于稳态属性的工业控制协议操作字段识别方法。首先,通过会话还原、碎片化包重组等预处理操作,从会话数据中提取出数据包应用层各个字段的取值序列;然后,通过对各个取值序列稳定性、周期性、相关性进行分析,发现操作字段的取值序列具有较为稳定、高周期性、高相关性的稳态属性;最后,通过无监督聚类的方式对操作字段和其他字段进行了有效划分,实现操作字段的自动识别。在多种工控系统环境下进行了广泛验证,测试数据包括电力网、水处理实验平台数据以及实网工控流量数据。结果表明,操作字段的识别率在90%以上,证明了方法的有效性和通用性。

关键词:工业控制协议;操作字段;稳态属性;字段识别

中图分类号:TP38　　**文献标志码:**A　　**文章编号:**2096-3246(2025)05-0355-12

根据美国国家标准技术研究院(NIST)给出的定义[1],工业控制系统是一个涵盖了多种类型的控制系统通用术语,包括监控和数据采集(SCADA)系统、集散控制系统(DCS)等。工业控制系统常用于工业领域和关键基础设施的控制中,其生产控制逻辑通过网络操作指令来执行,数据包作为通信流量的组成单位承载了丰富的操作指令。通过数据包应用层字段的取值可以了解每一个操作的具体信息,如操作对象(如设备、组件、功能单元等)、操作类型(如读取、写入、控制等)及更加详细的操作参数(如读取或写入的数量等),是学习系统行为运行特征、分析业务特点的关键窗口[2]。然而,工控协议应用层中除包含标识操作的字段之外,还包含了其他类型的字段(如序号、身份标识(ID)号、校验码等)及返回的现场状态数据。操作字段识别的是否全面、准确将直接影响行为模式构建、业务态势分析、异常行为检测等任务的效果。

目前,针对操作字段识别主要有基于协议规范和基于程序执行两种方式[3]。因为公开的工控协议(如Modbus、S7、CIP协议[4–5]等)具有明确的协议规范,对协议应用层各个字段的语义有明确定义,所以通过分析协议规范的方式实现操作字段识别是最初的研究思路:如Ghosh等[6]通过人工分析Modbus协议各个字段的作用,识别了其中的功能码等字段,通过操作字段熵值完成对Dos攻击和中间人攻击(MITM)[7]的检测。而对于不同的工控协议,协议应用层中操作字段的长度、数量、位置都不相同,需要根据相应的协议规范进行分析识别,如马标等[8]对S7协议进行语义分析,识别出协议中ROSCTR、Parameter Length等操作字段来定义状态事件。Tian等[9]采用深度解析算法识别了CIP协议的Command等操作字段用于构建白名单。其余研究[10–13]则是针对IEC–60870–5–104、IEC61850等不同协议,通过人工选取的方式识别出相应的操作字段。上述识别方法虽然能够根据协议规范完成对操作字段的精准识别,但识别结果对工控系统

收稿日期:2023–12–13　　**修回日期:**2024–03–13　　**网络出版日期:**2024–06–03

基金项目:2020年工业互联网创新发展工程项目(TC200H01V)

作者简介:覃　朗(1993—),男,硕士.研究方向:网络威胁检测.E-mail:2021223045168@stu.scu.edu.cn

* **通信作者:**陈兴蜀,教授,E-mail:chenxsh@scu.edu.cn

内部采用的协议具有极强的依赖性，识别方法通用性较低。此外，该类方法也无法处理未公开协议细节的私有协议。

基于程序执行的识别方法通过跟踪并分析目标工控协议程序字段定义、函数调用、操作序列、内存数据、寄存器信息等方式，完成对操作字段的识别。如通过使用软件 IDA Pro 获取工控程序静态反汇编代码来识别操作字段[14]或通过调试、代码注入等手段对工控程序进行分析，并结合产生的流量完成字段语义的推断与操作字段的识别[15]。但此类方法仍然依赖大量的人工操作分析。

此外，在与本文研究目标类似的协议逆向分析领域，对于协议中的一般字段的识别分类问题，研究者提出了多种解决方法，例如基于序列比对[16]、频繁项挖掘[17]、概率推断[18−19]及深度学习[20]等，可以自动对字段的类型进行识别。然而，这些方法主要关注于对字段的分类和未知协议的格式提取，而并非操作字段的识别。

综上所述，在对操作字段识别的相关研究中，工控领域具有组件协议私有多样、场景复杂多变的特性，导致传统的基于协议规范和基于程序执行的识别方法分别存在识别方法场景依赖性强、识别成本高等问题，无法在场景、协议未知的情况下快速对操作字段进行识别。

为解决上述问题，需要摆脱对于协议规范或工控程序的依赖，从工控系统本身的特点进行分析：不同于不同信息网络呈现多变、复杂、主观意识强的特点，工控系统以物理定律和领域规范为运行时约束，以组件间的按需交互为主要运行模式，以动态平衡、周期循环作为其运行特征，针对特定的业务目标，具有固定的操作流程，受人为等随机因素影响较小[21]，无论工控系统内部采用何种工控协议，其操作行为会呈现出周期性强、重复出现等特征。而这些特征会在数据包层面反映在操作字段上，使不同协议的操作字段具备一种共有的稳态属性。基于此，本文提出了一种基于稳态属性的工控协议操作字段识别方法。首先，通过分析工控协议应用层各个字段取值在时间上的变化，提出了稳定性、周期性、相关性 3 个可以明显表现在操作字段上的特征；然后，通过无监督聚类的方式对操作字段和其他字段进行了有效划分，实现操作字段的自动识别。本文所提出方法的重要特征是它的多域适应性（对来自不同场景、不同协议的工控系统的网络流量使用相同的操作字段识别方法）。分别在电力网、水处理实验平台数据及实网工控流量数据下进行了验证，操作字段的识别率在 90% 以上。

本文的主要贡献如下：1）针对工控协议操作字段识别的通用性问题，提出了一种基于稳态属性的操作字段识别方法，通过操作字段呈现出的稳定性、周期性等稳态属性进行无监督聚类，实现了自动识别工控协议操作字段的目标；2）为解决部分工控协议因分段传输带来的报文碎片化的问题，结合工控协议传输特点，提出了 COTP–Rec 算法以完成数据包应用层数据的完整还原；3）使用不同协议场景下采集的流量数据评估本文所提的方法，在每个场景下都取得了 90% 以上的操作字段识别率，验证了本文所提方法的有效性和多域适应性。

1 字段特征分析

1.1 字段划分

本文的目标是在未知协议具体格式和语义定义的情况下实现操作字段的识别，无法利用先验知识得到各个字段的长度信息，也无法知道应用层头部字段与载荷的分界位置信息，所以本文将应用层的格式（包括头部字段和载荷）按照 1 个字节为单位进行统一划分，以 Modbus/TCP（Modbus 协议中一个应用最为广泛的变体，以下无特别说明，Modbus 协议都默认指的是 Modbus/TCP）协议为例，将其应用层前 10 个字节的格式进行划分，得到 Pos–1～Pos–10 共 10 个字段，Pos 后的数字代表字段或载荷在应用层中的位置，Modbus 协议应用层字段划分的结果如表 1 所示。

表 1 字段划分结果

Tab. 1 Field division result

应用层原格式	字段划分结果
Transaction ID	Pos–1、Pos–2
Protocol ID	Pos–3、Pos–4
Length	Pos–5、Pos–6
Unit ID	Pos–7
Func Code	Pos–8
Count	Pos–9
Payload	Pos–10

表 1 中：Transaction ID 标识了数据包的序号；Protocol ID 是一个固定字段；Length 字段标识了该字段之后数据包的长度；而 Unit ID 字段标识了从站地址，即操作对象；FuncCode 字段标识了操作类型；Count 标识了操作的数量（读取或写入几个字节），即操作参数；Payload 则代表返回的状态数据（这里仅显示了 1 个字节的状态数据）。

1.2 碎片化包重组

在完成字段划分的基础上，如果应用层数据完

整,理论上可以通过字段在应用层的位置得到相应的取值。但部分的工控协议(如S7、IEC–61850协议等)会出现应用层数据分段传输的现象,造成流量中出现大量碎片化的数据包,使得字段取值错误,影响后续字段特征分析的工作。

不同于常见的TCP分段和IP分片,工控协议的分段传输是通过COTP[22]协议来实现的,COTP是位于开放系统互连(OSI)标准模型传输层与应用层之间的一类,为上层的应用层协议提供数据传输服务,COTP有多种类型,如连接确认与断开类型,数据传输(DT)类型等。COTP数据传输类型的协议格式如图1所示,其中,Length标识了COTP的长度,PDU Type标识了COTP的数据包类型,TPDU Number标识了传输序号,LastData Unit则标识了是否为最后一个分段,Data为分段传输的数据,在本文中特指工控协议应用层的数据。

Length	Type	TPDU Number	LastData Unit	Data

图1　COTP数据传输数据包格式

Fig. 1　COTP data transmission packet format

为解决COTP分段传输带来的数据包应用层碎片化的问题,本文提出了COTP–Rec算法对分段数据进行重组,以还原出完整的应用层数据。COTP–Rec首先定义了两个缓冲区Buffer1和Buffer2来暂时存储上行流和下行流的分段数据,然后通过LastData Unit字来判断是否为最后一个分段,如果不是,则将分段数据缓存至对应缓冲区,如果是,则将缓冲区内数据与分段数据拼接,最终形成新的完整的应用层数据。COTP–Rec的算法代码如下所示:

```
输入: input_pcap_file                #输入待重组pcap
输出: output_pcap_file               #输出重组完成pcap
1.with PcapReader(input_pcap_file) as reader,
2.PcapWriter(output_pcap_file) as writer:
3.    current_cotp_data_ctos = b""
4.    # 用于缓存从客户端到服务端的当前 COTP 数据段
5.    current_cotp_data_stoc = b""
6.    # 用于缓存从服务端到客户端的当前 COTP 数据段
7.    for packet in reader:
8.       if Raw in packet:
9.          raw_data = bytes(packet[Raw])
10.         pdu_type = raw_data[5]  #获取COTP类型字段
11.         if (pdu_type != TYPE_DT):
12.            continue
13.         last_data_unit = raw_data[6] # 获取 lastdataunit
14.         if last_data_unit == LAST_SEGMENT:
15.            current_cotp_data_stoc += raw_data[7:] or
16.            # 根据包方向将当前分段写入对应缓存
17.            current_cotp_data_ctos += raw_data[7:]
18.            recdata += current_cotp_data_stoc
19.            # 重组缓冲区数据
20.            writer.write(Raw(recdata))
21.            # 输出该数据包到pcap
22.            current_cotp_data_stoc = b""
23.            or current_cotp_data_ctos = b"" # 清空缓存
24.         else:
25.            current_cotp_data_ctos += raw_data[7:]
26.            current_cotp_data_stoc += raw_data[7:]
```

1.3　字段取值序列提取

为了便于对字段取值变化进行分析研究,依次提取会话流的数据包中该字段的取值,形成字段取值序列。

会话流。从TCP报文3次握手开始建立连接到TCP报文4次挥手正常关闭连接或通过RST(reset)报文异常断开连接之间,具有相同5元组(源IP地址、源端口、传输层协议、目标IP地址、目标端口)和反转5元组(交换源和目标IP地址及端口)的全部数据包的集合构成一条会话流F_s,表示为:

$$F_s=<p_1,p_2,\ldots,p_i,\ldots,p_n> \tag{1}$$

式中,n为会话流内数据包的个数,p_i为会话流中第i个数据包(i=1,2,⋯,n)。

字段取值序列S为会话流中数据包该字段的取值按照时间先后顺序排列而成的数列:

$$S=<v_1,v_2,\ldots,v_i,\ldots,v_n> \tag{2}$$

式中,v_i表示该字段在第i个数据包p_i中的取值。

以Lemay SCADA数据集[23]为例,对Modbus各个字段的取值序列进行提取。Lemay SCADA数据集通过SCADA沙箱[24]生成仿真Modbus流量,并使用电网模拟器在流量中引入了真实性。图2为Pos–1～Pos–10的字段取值序列。

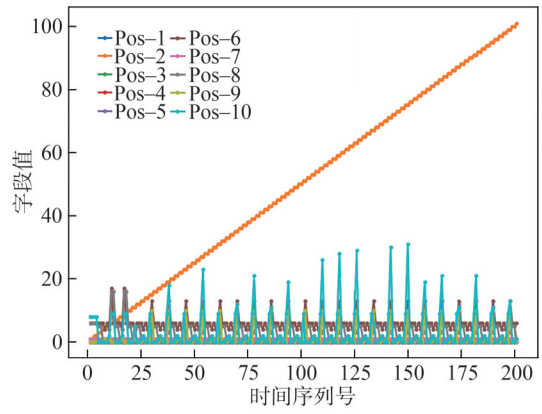


图2　Modbus Pos–1到Pos–10的取值序列

Fig. 2　Value sequence of Pos–1 to Pos–10

1.4 字段特征分析

1.4.1 稳定性

不同的工控协议中包含了不同种类和数量的操作类型(如 Modbus 协议的读取线圈、写入寄存器等，S7 协议的写入值、读取值等)。图 3 为不同工控协议常见操作类型的数量。

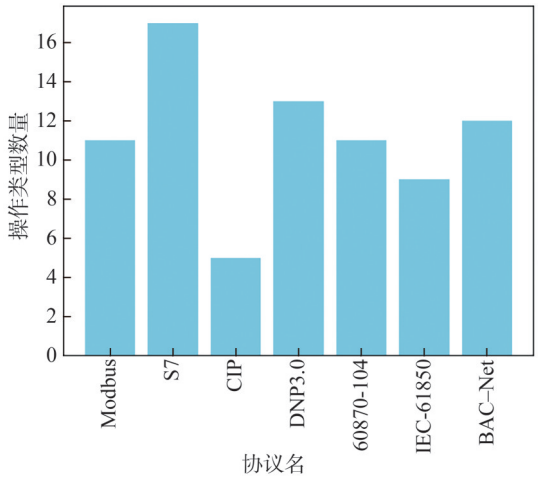


图3 不同协议的操作类型数量

Fig. 3 Number of operation types for different protocols

从图 3 中可以看出，虽然不同协议的可选操作类型数量不尽相同，但是数量取值都位于 5～16 这一较小的范围之内，而操作类型是由操作字段的取值所决定的，所以操作字段的可选取值数量从理论上也是较为有限的，应具备稳定性这一特征。

为了便于分析每一个字段的稳定性特征，将原有取值序列按照一定大小的时间窗口进行分段，计算每个段内子取值序列的熵值，得到分段统计熵值后的取值序列 S_E：

$$S_E=<E_1,E_2,\cdots,E_n> \tag{3}$$

式中，E_n 表示分段后的统计熵值，表示在时间窗口内字段取值的稳定性。图 4 为经过分段熵值计算(时间窗口=256)后的各个字段的取值序列。

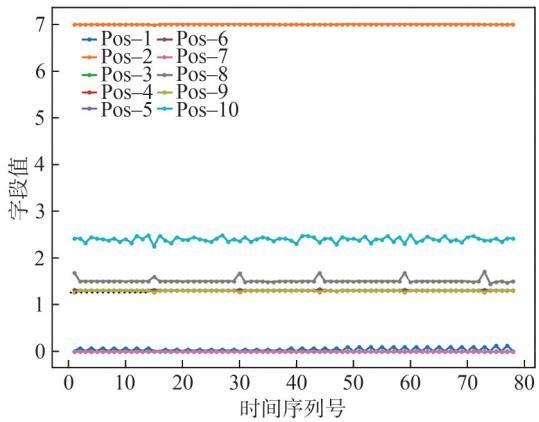


图4 各字段取值序列(熵值统计)

Fig. 4 Value sequences of each field (entropy statistics)

从图 4 可以看出，各个字段的熵值分布存在差异性，即字段的稳定性存在差异，表现为：不同的字段对应的熵值取值序列分布区域不同。具体大致可以分为 4 类。A1：Pos–3、Pos–4、Pos–7 字段，其熵值为 0，表明字段取值固定。A2：Pos–1、Pos–5 字段，其值取值序列中的熵值趋近于 0 但不都等于 0，说明这类字段的取值极少发生变化，但又不完全固定。A3：Pos–2 字段，其熵值趋近于 7，这说明在一个时间窗口内，其不同取值的数量较多。A4：Pos–6、Pos–8、Pos–9、Pos–10 字段其熵值变化位于 1.32～2.40 之间。

上述分类中的 A1、A2、A3 的稳定性呈现出过高或过低的特点，说明字段取值偏向于固定或随机的特性。而 A4 中的字段的熵值都位于一个较小的范围之内，进一步对照表 1 中对 Modbus 协议的字段划分情况可知，A4 中的 Pos–6、Pos–8、Pos–9 分别对应属性 Length、FuncCode 及 Count，分别标识了操作长度、操作类型及操作数量，属于 Modbus 协议中的操作字段，说明为了确保工业控制的稳定性和可靠性，操作字段在一定时间范围内的变化相对稳定，不会出现过大的波动和不确定性。

1.4.2 周期性

从上述按照字段熵值分布情况进行分类的结果可以得知，操作字段都属于第 A4 类字段，在对操作字段识别过程中，可以快速过滤掉 A1～A3 这几类熵值过高或过低的字段，但 A4 中除了包含操作字段之外，还包含了载荷。操作字段和载荷之间的熵值表现差异并不明显(载荷熵值稍高于操作字段)，在稳定性特征上存在重叠，需要增加分析特征的维度。工控系统控制操作具有循环、周期性强的特点，而这些特点可能会在数据包层面反应到操作字段上。所以本小节对各个字段的周期性进行分析。

本文使用自相关系数来描述一个取值序列的周期性特征，通过对字段取值序列 S 延迟不同的阶数 x (x=1，2，…，n)后得到延迟序列集合 S^x，将 S^x 与 S 进行相关性计算，得到一组自相关系数形成的有序集合 L_{SelfCorr}。

$$L_{\text{SelfCorr}}=<\rho_{S,S^1},\rho_{S,S^2},\cdots,\rho_{S,S^n}> \tag{4}$$

式中，ρ_{S,S^x} 为原序列 S 和延迟序列 S^x 的皮尔逊相关系数，即两者的协方差与标准差 s_{td} 的商。

Pos–3、Pos–4、Pos–5、Pos–7 所对应的取值序列是恒定值，故不存在自相关系数图，分别绘制其他字段(Pos–1、2、6、8、9、10)取值序列对应的自相关系数图，如图 5 所示。

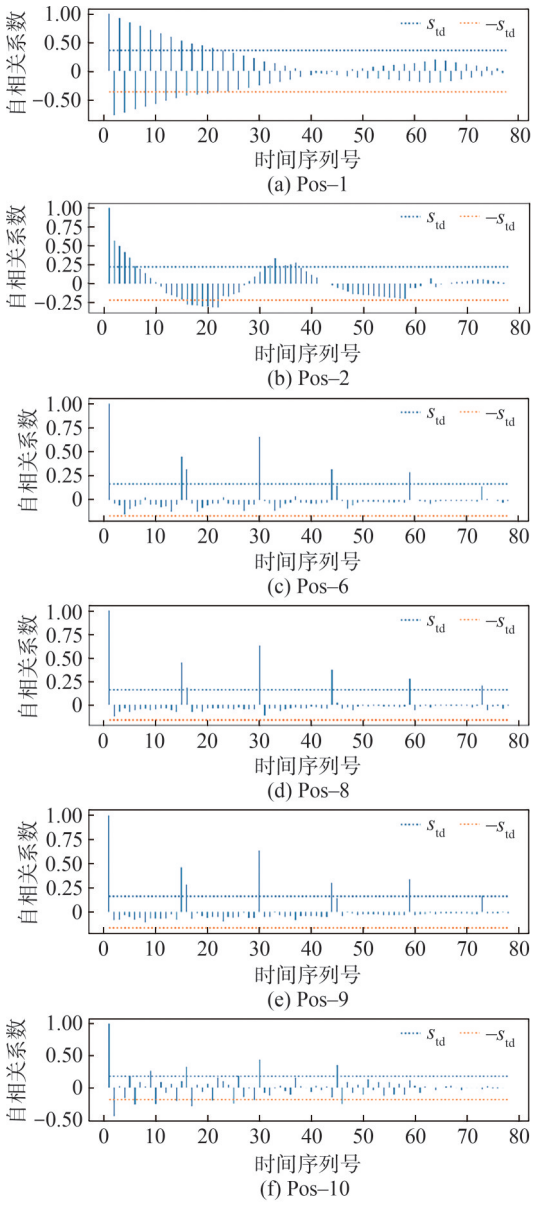


图5　各字段取值序列自相关系数

Fig. 5　Autocorrelation coefficients of each field value sequence

从图5和自相关系数结果可以看出，取值序列的周期性特征有以下4类。B1：由于Pos–3、Pos–4、Pos–5、Pos–7取值恒定，不存在自相关系数图，其周期可为任意大于0的数值。B2：Pos–1、Pos–2的自相关系数图中，自相关系数存在显著非0的情况，说明其周期性特征较为明显。但是自相关系数衰减为小值波动的过程比较缓慢且连续，呈振荡衰减的形式，具有明显的拖尾特征。B3：Pos–6、Pos–8、Pos–9的自相关系数图同样存在显著非0情况，但与B2不同的是自相关系数变化具有明显的截尾特征。B4：Pos–10的自相关系数图不具备明显的拖尾和截尾特征，周期特性亦不明显，自相关系数整体呈较低水平。

上述分类中的B3对应于Modbus的操作字段，具有较高的自相关系数，周期性特征较强。这说明为完成工控系统周期性的控制逻辑，操作字段会在周期内以固定的取值模式进行循环，具备较强的周期特征。而B4中的Pos–10对应的载荷则不具备明显的周期特征，可以与操作字段很好地区分开来。

1.4.3　相关性

根据周期性对字段的分类结果，可以快速地过滤掉不具备明显周期性的字段（B1和B4），但B2和B3都具备较强的周期性，其中B2对应于Modbus的Transaction ID字段，属于序号类型字段，其周期性是序号类字段递增的特性所导致的。所以，序号类字段和操作字段会因在周期性特征上产生重叠而变得难以区分。图6为操作字段Pos–6、Pos–8、Pos–9的字段取值序列图（熵值统计），可以发现各个取值序列在时间序列号为15、30、43、59、74时都发生了跳变，这说明各个操作字段的取值具有一定的相关性，可以作为识别控制字段的特征之一。

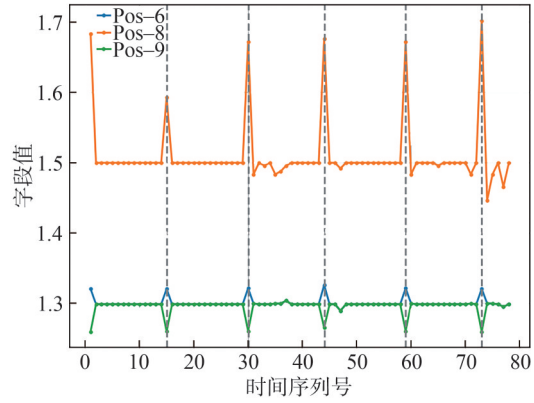


图6　Pos–6、Pos–8、Pos–9取值序列（熵值统计）

Fig. 6　Pos–6, Pos–8, Pos–9 value sequences

本文使用皮尔逊相关系数来衡量各个字段的相关性特征。皮尔逊相关系数是一种用于度量两个变量之间线性关系强度和方向的统计指标。通过皮尔逊相关系数计算字段两两之间的相关系数，构成如图7所示相关系数矩阵。

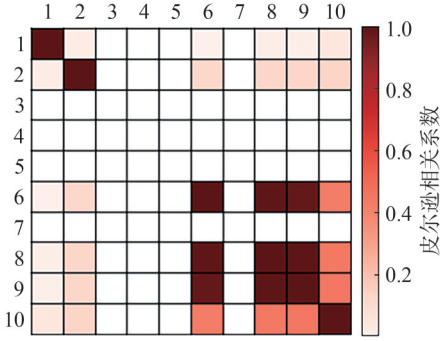


图7　字段取值的相关系数矩阵

Fig. 7　Correlation coefficient matrix of field values

图7中，Pos–6与Pos–8、Pos–6与Pos–9、Pos–8与

Pos–9 所对应的区块都呈现出深色，证明具有强相关特性，这说明操作字段满足相关性特征条件，即操作字段的取值通常会与其他操作字段的取值存在一定关联关系。而其他字段的相关性并不高，表明其取值相对独立。

1.5 操作字段的稳态属性

通过上述对各个字段稳定性、周期性、及相关性的分析，可以得知，其他类型字段和载荷不具备或仅具有上述特征的 1 个（如序号类字段具备周期性，载荷具备稳定性等），而仅有操作字段同时具备如下属性。

1）较为稳定：因工控协议操作字段的可选取值相对较少，其稳定性表现小于固定字段而大于序号类字段。

2）强周期性：因工控系统本身的逻辑控制操作具有强周期性，所以操作字段具有较强的周期特征。

3）强相关性：操作是由操作对象、操作类型、操作参数所组成的统一整体，对于特定的操作对象有相对应的操作类型和操纵参数。所以，操作字段之间存在着较强的相关性。

上述属性既与场景中采用协议的类型无关，也与使用的工控程序无关，是操作字段所具备的跨场景、跨协议的稳态属性。

2 操作字段识别

为了完成对操作字段的自动识别，需要首先对字段的各个特征进行量化表述，分别定义稳定性系数、周期性系数及相关性系数以衡量字段 3 类特征的强弱。

2.1 特征属性量化

1）稳定性系数 C_{sta}

C_{sta} 表示字段的稳定性程度。定义一个可接受的熵值范围区间 $[E_{\min},E_{\max}]$，$E_{\min}$ 和 $E_{\max}$ 根据经验分别取一个较保守的值 1 和 6。当字段的熵值位于这个区间范围之外，则表明字段的稳定性过高或过低，直接使稳定性系数为 0。当位于这个区间范围内，则需要通过计算熵值的均值和标准差来得到稳定性系数。具体地，有如下计算公式：

$$C_{\text{sta}}=\begin{cases}0, S_E\notin[E_{\min},E_{\max}];\\ 1/(\mu\sigma), S_E\in[E_{\min},E_{\max}]\end{cases}\tag{5}$$

式中，μ、σ 分别为熵值的均值和标准差。

2）周期性系数 C_{per}

周期性系数 C_{per} 反映字段周期性的强弱程度，这里取自相关系数里的最大值作为周期性系数，记为：

$$C_{\text{per}}=\max(L_{\text{SelfCorr}})\tag{6}$$

3）相关性系数 C_{corr}

相关性系数 C_{corr} 反映字段相关性的强度，定义 C_{corr} 为字段 j 与其余所有字段（个数为 n，索引为 i）的皮尔逊相关系数之和，记为：

$$C_{\text{corr}}=\sum_{i(i\neq j)}^{n}\rho_{i,j}\tag{7}$$

按照上述方法对各系数进行计算，并进行归一化操作后，每一个字段都对应了一个由稳定性系数、周期性系数、相关性系数组成的 3 维特征空间：

$$C=<C_{\text{sta}},C_{\text{per}},C_{\text{corr}}>\tag{8}$$

2.2 字段聚类

根据第 1.4 节的字段特征分析结果，仅有操作字段同时具备较高稳定性、较强周期性、相关性的稳态属性。基于此将每个字段对应的特征空间输入到算法模型中进行无监督的聚类，以完成操作字段的识别。图 8 为 Pos–1 到 Pos–10 字段的聚类结果（聚类个数设定为 4）。

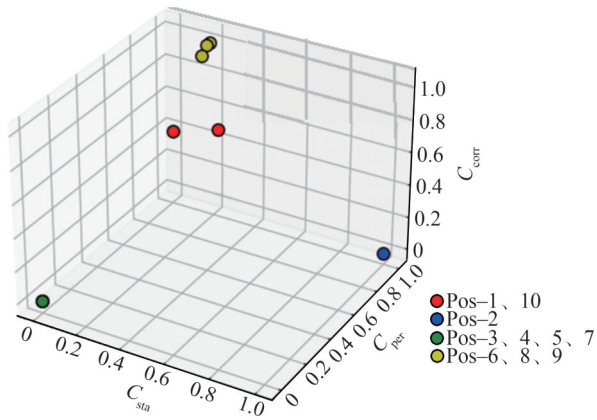


图 8 字段聚类结果

Fig. 8 Field clustering results

图 8 中，属于绿色类别的字段有 Pos–3、Pos–4、Pos–5、Pos–7，具有高稳定性，低周期性、相关性的特点。分别对应 Modbus 中的 Protocol ID（Pos–3，Pos–4）及 Length（Pos–5）和 Unit ID。因 Protocol ID 是取值不变的固定字段，而 Length 高位和 Unit ID 字段因为在实验数据的环境中并未使用，所以其各项特征和固定字段一样。

图 8 中，属于蓝色类别的字段有 Pos–2，具有高周期性、低稳定性、低相关性的特点。对应于 Modbus 中 Transaction ID 低位。Transaction ID 标识了数据包的序号，其值从 0 到 255 循环变化。

图 8 中，属于黄色类别的字段有 Pos–6、Pos–8、Pos–9，同时具备较为稳定、周期性和相关特征强的特点，分别对应 Modbus 中 Length 字段的低位、FuncCode 以及 Count 字段，都属于 Modbus 的操作字段。

图 8 中，属于红色类别的字段有 Pos–1、Pos–10，具有一定的稳定性以及周期性和相关性，但皆不明显，且与上述类别中各个字段分布集中不同，该类字段分布较为分散，分别对应于序号字段的高位和载荷。

3　实验与分析

3.1　实验数据

研究旨在实现在不同场景、不同协议下对操作字段的准确、自动提取。除 Lemay SCADA 数据集外,新增两个场景的真实流量数据评估所提方法的准确性和通用性:1)iTrust,新加坡科技设计大学网络安全研究中心,拥有多个世界级的测试平台和培训平台,通常会连续数天不间断地运行,收集数据用于研究人员使用,致力于提高传统和新的关键基础设施的安全性,考虑本文对数据要求为原始 pcap 文件,选择其中安全水处理(SWaT)[25]的数据集用于实验验证;2)除此之外,还有与作者所在团队实验室合作单位采集的真实世界中工控系统网络数据集为实验数据。

因为本文所提方法需要分析周期性等特征,对于会话的长度有一定的要求,所以对数据集中短会话进行了删除,仅保留会话长度大于 2 000 的长会话。实验数据的具体情况如表 2 所示。

表 2　实验数据集

Tab. 2　Experimental data set

名字	使用协议	数据集描述	数据集规模
Lemay SCADA	ModBus	SCADA 沙箱 10 h 仿真流量	会话数量 242,数据包数量 1 667 815
SWaT	CIP	安全水处理平台 1 d 的实验流量数据	会话数量 647,数据包数量 6 522 679
Real ICS	S7	真实电网系统中 3 h 流量数据	会话数量 712,数据包数量 3 223 143

3.2　评估指标

准确性是识别方法及模型评估的核心,常见的准确性评价指标有两个:真阳性率(TPR)和假阳性率(FPR)。由于操作字段识别的特殊性,结合真阳性率和假阳性率的思想,本文引入识别率(TIR,记为 T_{IR})和误识别率(FIR,记为 F_{IR})两个指标来对算法准确性进行的评估。

$$T_{IR}=\frac{\text{正确识别出的操作字段数量}}{\text{实际的操作字段数量}} \tag{10}$$

$$F_{IR}=\frac{\text{将非操作字段识别为操作字段数量}}{\text{数据包应用层长度}} \tag{11}$$

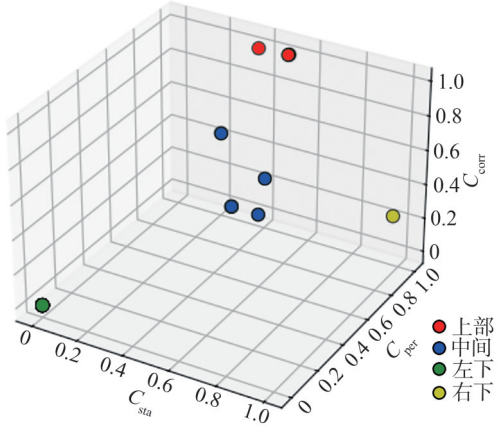
3.3　聚类算法选择

本文对比了多种聚类算法的效果,包括基于距离的 K–means、基于密度的 DBSCAN、基于概率分布的软聚类方法—高斯混合模型(GMM)以及基于相似度的谱聚类算法[26]。其中,K–means 聚类结果的平均值有最高的识别率以及较低的误识别率,所以本文采用 K–means 算法将字段划分为 K 个簇(K 值经实验验证最佳值为 4)。

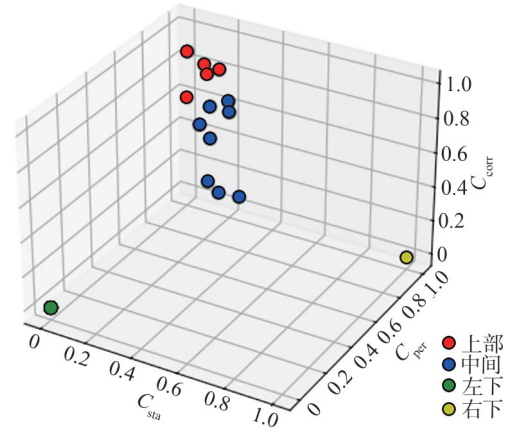
3.4　实验结果

3.4.1　聚类结果

图 9 为 SWAT 和 Real ICS 数据集字体聚类结果。从 SWAT(图 9(a))、Real ICS(图 9(b))和 Lemay SCADA(图 8)的字段聚类结果来看,三者聚类簇的分布具有相似性,分别位于 3 维特征空间的左下角、右下角、上部和中间位置。其中,左下角聚类簇代表仅具有高稳定性的字段,右下角代表仅具有强周期性,而上部簇是具有较高稳定性、强周期性、强相关性,即稳态属性的字段。而中部簇则是 3 个特征皆不明显的字段。上部聚类簇内的所有字段即为本方法的识别结果。



(a) SWAT聚类结果



(b) Real ICS聚类结果

图 9　SWAT 和 Real ICS 数据集字段聚类结果

Fig. 9　SWAT and Real ICS dataset field clustering results

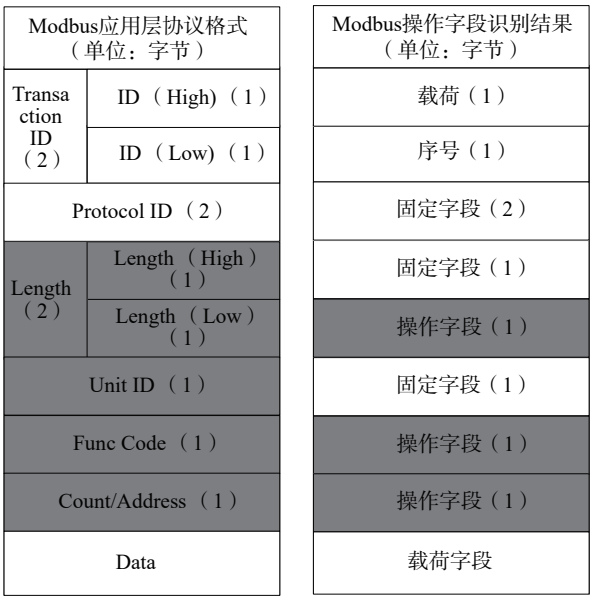
3.4.2　识别结果

为了对识别结果进行验证,首先分别依据 Modbus、S7、CIP 协议规范中对各个字段的作用定义来标注字段是否为操作字段;然后,与本文所提方法的识别结果进行对比分析。

1)各协议操作字段识别结果分析

图 10 为 Modubs 操作字段识别结果,其中,()中的数字为字段所占字节数。图 10(a)为 Modubs 应用层协

议格式(操作字段区域标注为深色,下文同),10(b)为针对一条 Modbus 会话的操作字段识别结果。对比图 10(a)和(b)可知,方法正确识别出了 Modbus 协议中 Length 字段的低位、功能码以及 Count 字段为操作字段,但将 Length 字段的高位识别为固定字段。因为该会话中,数据包长度较小,所以 Length 字段的高位并未使用;但 Unit ID(从站地址)在 Modbus/TCP 中被 IP 地址所取代,所以也被方法识别为固定字段。



(a) 协议格式 (b) 识别结果

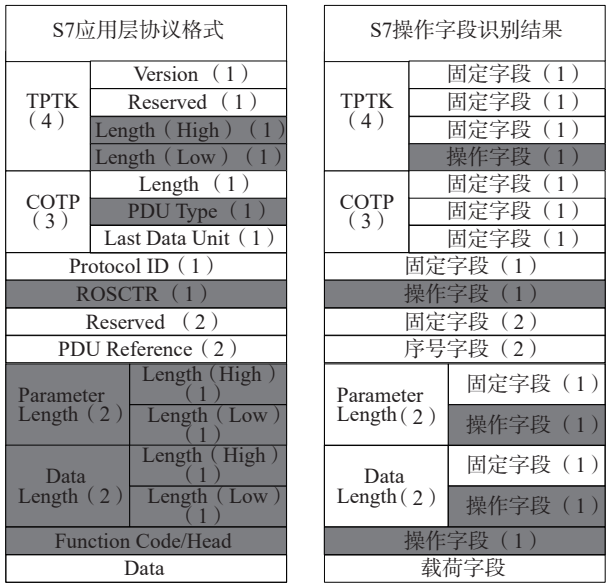
图 10 Modbus 操作字段识别结果

Fig. 10 Modbus operation field identification results

图 11 为 S7 应用层协议格式(图 11(a))和针对一条 S7 会话的操作字段识别结果(图 11(b)),对比图 11(a)和(b)可知,方法正确识别出了 S7 协议中 TPTK–Length 字段的低位、ROSCTR(操作类型字段)、Parameter–Length 及 Data–Length 的低位以及 Function Code 字段为操作字段,其余未识别出字段的原因同 Modbus 类似,在此不再赘述。

图 12 为 CIP 应用层协议格式(图 12(a))和针对一条 CIP 会话的操作字段识别结果(图 12(b)),对比图 12(a)和(b)可知,方法正确识别出了 CIP 协议中 Command 和 Length 字段的低位、Tieoutout 字段(操作超时时间)、Item Count(操作数量)、Command 字段及 Function Code 字段为操作字段。特别地,方法并未正确识别出 Services 字段为操作字段,原因是 Services 内的操作字段单位为比特级,但方法的识别精度仅为字节级。

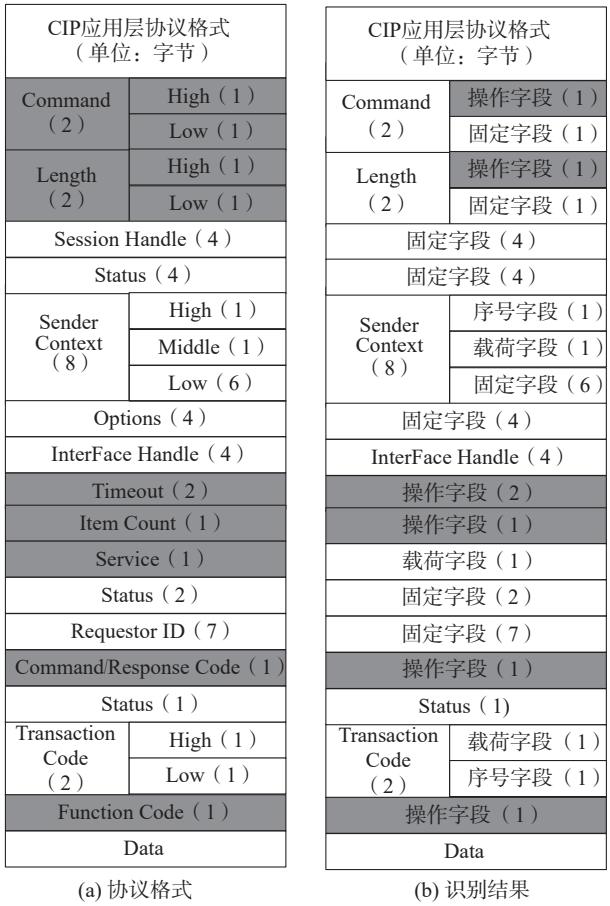
综合对 Modbus、S7 协议及 CIP 协议的识别结果来看,本文所提方法能够对大部分操作字段进行正确识别,对于某些在会话过程中未使用的操作字段也可以正确地进行排除,即识别结果的动态性,这也是相较于传统的基于协议规范进行操作字段识别方法所具备的独特优势。但本方法设计的识别精度为字节级,对于比特级操作字段识别效果可能不佳。



(a) 协议格式 (b) 识别结果

图 11 S7 操作字段识别结果

Fig. 11 S7 operation field identification results



(a) 协议格式 (b) 识别结果

图 12 CIP 操作字段识别结果

Fig. 12 CIP operation field identification results

2)识别综合效果评估

为了评估本文所提方法的识别效果,选择协议逆向分析领域中较为先进的方法 Netplier[18]进行对比。Netplier是一种基于概率推断的协议逆向分析方法,它通过对静态的数据包进行多重序列比对和关键字推断,进而对消息进行聚类,完成格式提取、语义推断和状态机重构等工作。尽管 Netplier 没有直接识别出操作字段,但可以将在关键字推断阶段得到的 FD(format distingguisher)字段(用于标识数据包类型的字段,Netplier 依据此字段的取值对数据包进行聚类)和 Length 字段作为 Netplier 的操作字段识别结果,以便与本文所提方法进行比较。

a. 数据集大小对方法的影响。在实际环境中,可能会出现数据集样本不充足的情况,所以对第 3.1 节中提到的实验数据按照 10%、30%、50%、100%进行随机采样,使用不同的数据集大小对本文所提方法与 Netplier 的识别效果进行评估,如图 13 所示为不同数据集大小情况下的识别率和误识别率。

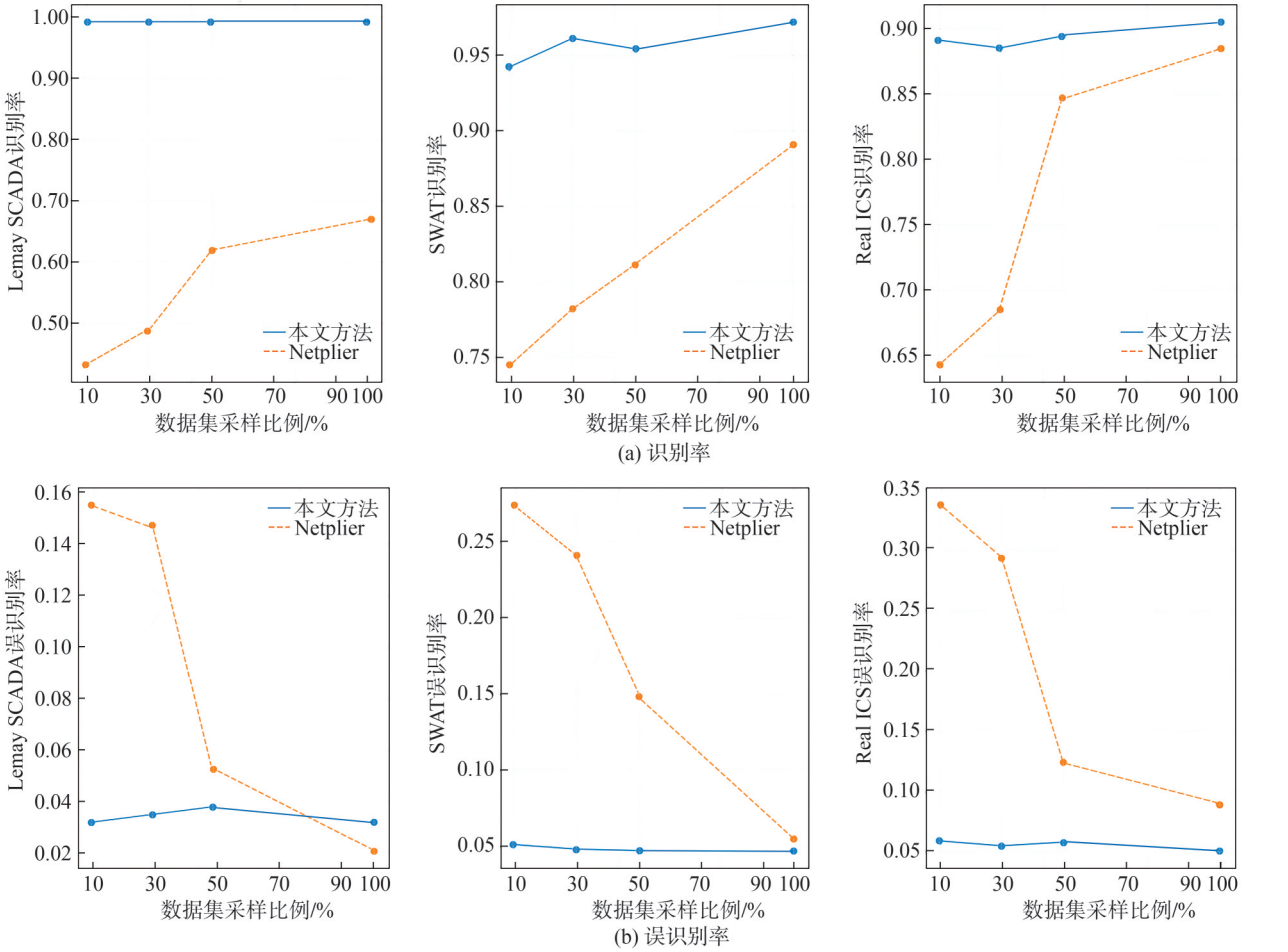


图 13　数据集大小对方法识别效果的影响

Fig. 13　Effects of dataset size on the recognition performance of methods

从图 13 中可以看出,在数据集大小为 100%时,本文所提方法在 3 个数据集上的识别率分别达到了 99.9%(Lemay SCADA)、97.2%(SWAT)和 90.5%(RealICS),高于 Netplier 的 67.7%(Lemay SCADA)、89.1%(SWAT)和 88.5%(RealICS),误识别率在 3 个数据集上都低于 5%。并且随着数据集的减小,Netplier 的识别率迅速下降,而误识别率迅速上升,因为 Netplier 识别前提是数据集中有足够多格式的数据包,才能完成后续的聚类分析等工作。而本文所提方法的效果基本不受数据集大小影响,对数据集 Lemay SCADA 和 SWaT 识别率稳定维持在 94.0%以上,Real ICS 场景下识别率也维持在了 88.0%以上。证明本文所提方法可以在数据量较小的情况下对操作字段进行正确识别。

b. 数据集质量对方法的影响。第 3.1 节中提到的实验数据在理论上是干净的,这里的干净指的是无噪声、人工操作或遭受攻击,而实际环境中数据集可能遭受污染,所以为了验证数据集质量对方法的影响,在数据集各个会话中随机插入 0、1.0%、2.5%、5.0%比例的坏包,坏包由当前会话内的数据包随机选择,用于模拟数据被污染的情况。图 14 为不同数据集质量情况下的识别率和误识别率。

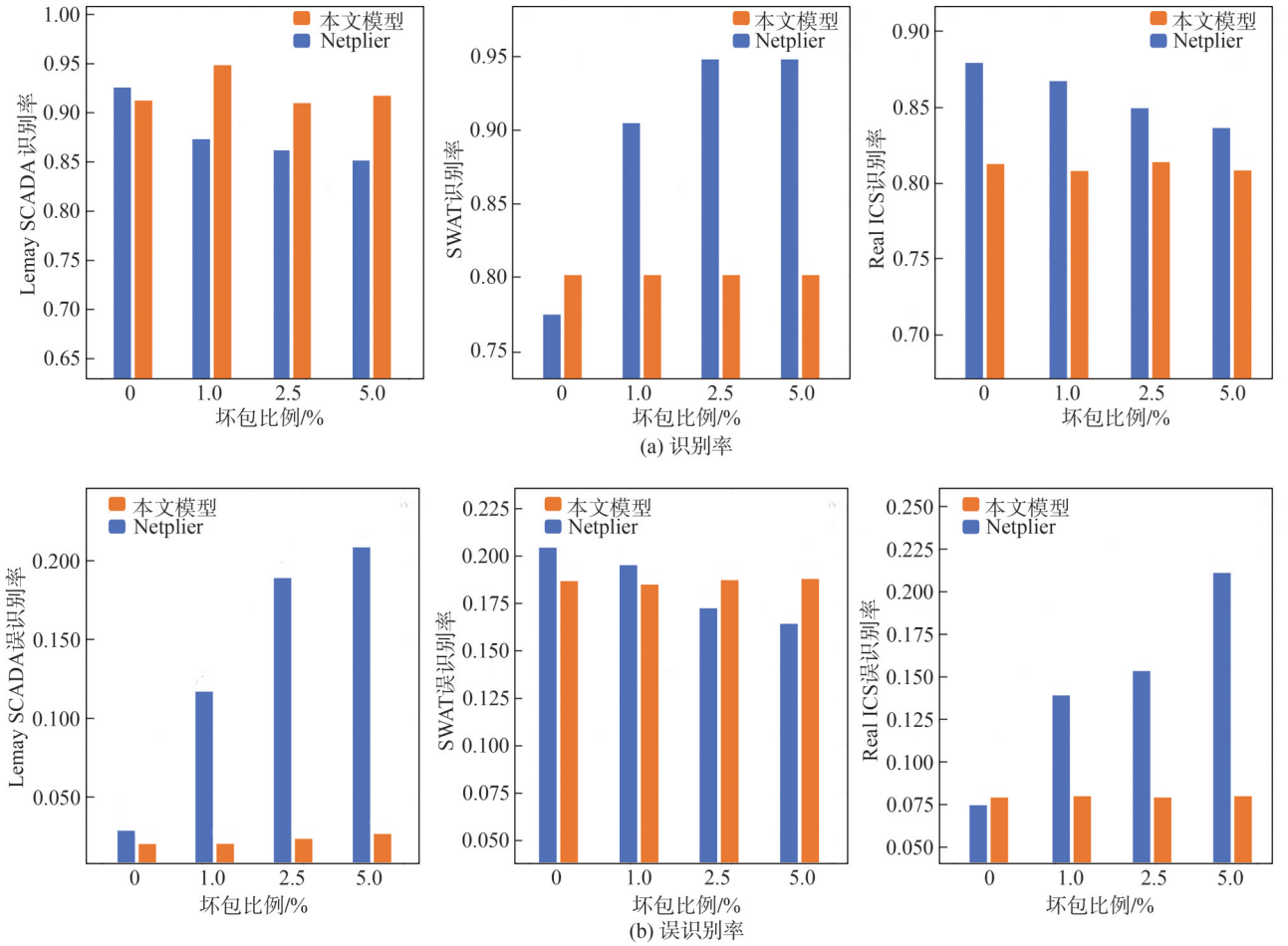


图 14　数据集质量对方法识别效果的影响

Fig. 14　Effects of dataset quality on the recognition performance of methods

从图 14 可以看出，数据集质量对于本文所提方法的效果影响较大，在 5.0% 坏包比例的情况下，Lemay SCADA 和 SWaT 识别率下降到 80% 以下，误识别率大于 20%，而 Real ICS 识别率在 67% 左右，误识别率在 25% 左右。这主要是因为坏包的添加破坏了操作字段周期性强、稳定等稳态属性，使得识别效果欠佳，这也是本方法的局限之一。而 Netplier 识别结果基本不受数据质量的影响，因为其分析的是静态的数据包，所以其对数据质量的要求不高。

4　结　论

本文针对现有操作字段识别方法依赖人工选取导致方法效率低下、通用性差的问题，提出了基于稳态属性的工业控制协议操作字段识别方法：首先，通过会话还原、碎片化包重组等预处理操作，从会话数据中提取出数据包应用层各个字段的取值序列；然后，通过对各个取值序列稳定性、周期性、相关性的分析，发现操作字段的取值序列具有较为稳定、高周期性、高相关性的稳态属性；最后，通过无监督聚类的方式对操作字段和其他字段进行了有效划分，实现操作字段的自动识别。分别在电力网、水处理实验平台数据以及实网工控流量数据下进行了验证，操作字段的识别率在 90% 以上，高于所选基线方法。本文所提方法在工控协议安全测试、工控行为监管、工控系统异常检测等方面都有较好的应用价值，例如，可以通过操作字段识别结果去生成有效的模糊测试数据。

本文所提方法能够在数据量较少的情况下完成识别任务，但对于流量的质量要求较高，即工控系统流量数据内不包含过多人为操作或噪音。此外，本方法的识别精度为字节级，对于比特级的操作字段识别效果可能不佳，这也是未来可以继续研究的方向。

参考文献：

[1] Stoffer K, Falco J, Scarfone K. Guide to industrial control systems (ICS) security[J]. NIST Special Publication, 2015, 800(82): 10–115.

[2] Hu Yan, Yang An, Li Hong, et al. A survey of intrusion detection on industrial control systems[J]. International Journal of Distributed Sensor Networks, 2018, 14(8): 155014771879461.

[3] Huang Tao, Fu Anmin, Ji Yukai, et al. Research and chal-

lenges on reverse analysis technology of industrial control protocol[J]. Journal of Computer Research and Development,2022,59(5):1015–1034.[黄涛,付安民,季宇凯,等.工控协议逆向分析技术研究与挑战[J].计算机研究与发展,2022,59(5):1015–1034.]

[4] Conti M, Donadel D, Turrin F. A survey on industrial control system testbeds and datasets for security research[J]. IEEE Communications Surveys & Tutorials, 2021, 23(4): 2248–2294.

[5] Ryalat M, ElMoaqet H, AlFaouri M. Design of a smart factory based on cyber-physical systems and Internet of Things towards industry 4.0[J]. Applied Sciences, 2023, 13 (4):2156.

[6] Ghosh T, Bagui S, Bagui S, et al. Anomaly detection for modbus over TCP in control systems using entropy and classification-based analysis[J].Journal of Cybersecurity and Privacy,2023,3(4):895–913.

[7] Bruschi D, Di Pasquale A, Lanzi A, et al. Ensuring cybersecurity for industrial networks: A solution for ARP-based MITM attacks[J].Journal of Computer Security,2024,32(5):447–475.

[8] Ma Biao,Hu Mengna,Zhang Chonghao,et al.Traffic anomaly detection method of industrial control network based on Fusion Markov Model[J]. Journal of Information Security, 2022,7(3):17–32.[马标,胡梦娜,张重豪,等.基于融合马尔科夫模型的工控网络流量异常检测方法[J].信息安全学报,2022,7(3):17–32.]

[9] Tian Zheng, Wu Weidong, Li Shu, et al. Industrial control intrusion detection model based on S7 protocol[C]//Proceedings of the 2019 IEEE 3rd Conference on Energy Internet and Energy System Integration.Changsha:IEEE,2019:2647–2652.

[10] Yang An, Sun Limin, Shi Zhiqiang, et al. Sbsd: Detecting the sequence attack through sensor data in ICSs[C]//Proceedings of the 2018 IEEE International Conference on Communications.Kansas:IEEE,2018:1–7.

[11] Gao Jianlei, Li Jun, Jiang Hao, et al. A new Detection Approach against attack/intrusion in Measurement and Control System with Fins protocol[C]//Proceedings of the 2020 Chinese Automation Congress.Shanghai:IEEE,2020:3691–3696.

[12] Wang Bin, Li Feng, Chen Tao, et al. Research on deep analysis technology of real time interaction protocol in power industrial control system[C]//Proceedings of the 2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence.Chongqing:IEEE,2021:75–80.

[13] Wang Bin, Zhang Jianye, Luo Cheng, et al. Research on deep detection technology of abnormal behavior of power industrial control system[C]//2022 IEEE 6th Information Technology and Mechatronics Engineering Conference. Chongqing:IEEE,2022,6:1256–1261.

[14] Wei Xiao, Liu Renhui, Xu Fengkai. Reverse analysis of industrial control protocol based on static binary analysis[J]. Application of Electronic Technique,2018,44(3):126–130.

[15] Ruan Wei, Huang Guangping, Chen Liang, et al. Deep analysis method of private protocol in industrial control system [J]. Electronic Technology & Software Engineering, 2019, 22:3–4.[阮伟,黄光平,陈亮,等.工业控制系统私有协议深度解析方法[J].电子技术与软件工程,2019,22:3–4.]

[16] Wu Zewei, Shu Min, Shi Junzheng, et al. How to reverse engineer ICS protocols using pair–HMM[M]//Information and Communication Technology for Intelligent Systems. Singapore:Springer Singapore,2018:115–125.

[17] Shim K S, Goo Y H, Lee M S, et al. Clustering method in protocol reverse engineering for industrial protocols[J]. International Journal of Network Management, 2020, 30(6): e2126.

[18] Ye Yapeng, Zhang Zhuo, Wang Fei, et al. NetPlier: Probabilistic network protocol reverse engineering from message traces[C]//Proceedings of 2021 Network and Distributed System Security Symposium. Chicago: The Internet Society,2021:24531.

[19] Chandler J, Wick A, Fisher K. BinaryInferno: A semantic-driven approach to field inference for binary message formats[C]//Proceedings of 30th Annual Network and Distributed System Security Symposium. San Diego: The Internet Society,2023:23131.

[20] Zhao Rui, Liu Zhaohui. Analysis of private industrial control protocol format based on LSTM–FCN model[C]//Proceedings of the 2020 International Conference on Aviation Safety and Information Technology. Weihai: ACM, 2020: 330–335.

[21] Narayanan S N, Joshi A, Bose R. ABATe: Automatic behavioral abstraction technique to detect anomalies in smart cyber-physical systems[J]. IEEE Transactions on Dependable and Secure Computing,2022,19(3):1673–1686.

[22] Liu Ran, Zhao Zhenyuan, Guan Zhiguang. Research on remote control system of excavator based on industrial Internet of Things[C]//Proceedings of the 13th International Conference on Computer Engineering and Networks. Singapore:Springer Nature Singapore,2024:492–500.

[23] Lemay A, Fernandez J M. Providing SCADA network data sets for intrusion detection research[C]//Proceedings of 9th Workshop on Cyber Security Experimentation and Test.Austin:USENIX Association,2016:6–6.

[24] Lemay A, Fernandez J, Knight S. An isolated virtual cluster for SCADA network security research[C]//Proceedings of 1st International Symposium for ICS & SCADA Cyber Security Research. Leicester: BCS Learning Development Ltd.,2013:88–96.

[25] Goh J, Adepu S, Junejo K N, et al. A dataset to support research in the design of secure water treatment systems [M]//Critical Information Infrastructures Security. Cham: Springer International Publishing,2017:88–99.

[26] Zhang Y, Zhou Y. Review of clustering algorithms[J]. Journal of Computer Applications,2019,39(7):1869.

Operation Field Recognition of Industrial Control Protocols Based on Steady-state Properties

QIN Lang[1,2], *CHEN XingShu*[2,3*], *ZHU Yi*[2,3], *LI Yao*[2,3], *HE Jun*[1]

(1.School of Computer Sciences, Sichuan University, Chengdu 610064, China;
2.School of Cyber Sciences and Engineering, Sichuan University, Chengdu 610065, China;
3.Cyber Science Research Institution, Sichuan University, Chengdu 610065, China)

Abstract: The operation fields in industrial control protocols play a critical role in recognizing industrial control network behavior, understanding and monitoring network activities, and accurately identifying and extracting operation fields from industrial control network traffic. However, current methods for operation field recognition often rely on expert experience or manual analysis based on program execution, resulting in low efficiency, limited generalizability, and an inability to handle many undisclosed proprietary protocols or automatically recognize operation fields in complex network scenarios with unknown contexts and protocols. Therefore, this study uses the unique domain characteristics of industrial control networks and proposes an operational field recognition method based on the steady-state properties of industrial control protocols, overcoming the limitations imposed by protocols and programs. First, by preprocessing industrial control network session data, such as session reconstruction and fragmented packet reassembly, the value sequences of various fields at the application layer of the data packets are extracted. Then, through analysis of the stability, periodicity, and correlation of these value sequences, operation fields exhibit steady-state properties characterized by stability, high periodicity, and high correlation. These steady-state properties are quantified as features of operation fields. Next, an unsupervised clustering method is employed to effectively distinguish operation fields from other fields, ultimately achieving automatic recognition of operation fields. The proposed method demonstrates significant value in industrial control protocol security testing, regulating industrial control behavior, and anomaly detection in industrial control systems. For example, by utilizing the recognition results of operation fields, it becomes possible to construct and generate effective fuzzy testing data to enhance the security of industrial control systems. Through extensive validation in various industrial control system environments, including power grids, water treatment experimental platforms, and real industrial control traffic data, the method achieves a recognition rate of over 90% for operation fields, demonstrating its effectiveness and generalizability. In addition, in the experimental section, the influence of data size and quality on the method is discussed in detail. The proposed method accomplishes the recognition task with relatively small amounts of data but requires high-quality traffic data with minimal artificial operations or noise in the industrial control system traffic. Therefore, in practical applications, it is important to ensure the accuracy and purity of industrial control system traffic data. In conclusion, the operation field recognition method based on the steady-state properties of industrial control protocols rapidly and accurately identifies operation fields by analyzing features such as stability, periodicity, and correlation, without relying on specific protocol specifications or source code analysis. The method provides essential technical support for industrial control network security monitoring and behavior analysis, while also providing new possibilities for intelligent control and management of industrial control systems.

Key words: industrial control protocols; operation fields; steady-state properties; field recognition

（编辑　吴芝明）

引用格式: Qin Lang,Chen XingShu,Zhu Yi,et al.Operation field recognition of industrial control protocols based on steady-state properties[J].Advanced Engineering Sciences,2025,57(5):355–366.[覃朗,陈兴蜀,朱毅,等.基于稳态属性的工业控制协议操作字段识别[J].工程科学与技术,2025,57(5):355–366.]