计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(7):236−242 [doi: 10.15888/j.cnki.csa.006443]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于语义向量与 OCSVM 的工控网络异常行为识别①

王佳楠[1,2], 李泽宇[3], 李喜旺[1]

[1](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[2](中国科学院大学, 北京 100049)
[3](国家电网公司 东北分部, 沈阳 110180)
通讯作者: 王佳楠, E-mail: wangjianan15@mails.ucas.ac.cn

摘　要: 为克服基于漏洞库等传统安全防护策略的短板, 实现对未知攻击行为的识别和预警. 使用时间窗划分和深度包检测技术, 将端到端的通信内容转化为控制行为序列. 根据工控协议的语义特性, 采用语义向量模型将行为序列转化为统一维度的特征向量. 基于单类支持向量机 (OCSVM) 仅使用正常行为样本构造的异常识别模型, 克服了无法从生产环境中获得异常样本的困难. 对于所仿真出的多种异常行为序列, 模型识别的平均准确率能够达到 93% 以上.

关键词: 工控网络; 语义向量; 特征提取; 单类支持向量机; 异常行为识别

引用格式: 王佳楠,李泽宇,李喜旺.基于语义向量与 OCSVM 的工控网络异常行为识别.计算机系统应用,2018,27(7):236–242. http://www.c-s-a.org.cn/1003-3254/6443.html

Identification of Abnormal Behavior in Industrial Network Based on Semantic Vector and OCSVM

WANG Jia-Nan[1,2], LI Ze-Yu[3], LI Xi-Wang[1]

[1](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](Northeast Branch Corporation, State Grid Corporation of China, Shenyang 110180, China)

Abstract: In order to overcome the shortcomings of the traditional security protection strategy based on the vulnerability database, the recognition and early warning of unknown attack behavior should be realized. Using time window division and deep packet inspection, the content of end-to-end communication is transformed into a sequence of control actions. According to the control protocol's semantic features, the control behavior sequences are transformed into the feature vectors of unified dimension using the semantic vector model. The anomaly recognition model based on One Class Support Vector Machine (OCSVM) is constructed by normal behavior samples only, overcoming the difficulty of obtaining exception samples from the production environment. The average recognition accuracy of the model is to more than 93% on the simulation sequences containing multiple abnormal behaviors.

Key words: industrial network; semantic vector; feature extraction; One Class Support Vector Machine (OCSVM); abnormal behavior recognition

在工业生产领域向网络化、信息化、自动化、拓展化的发展过程中, 大量的网络化控制设备和数据交换设施在提高工业生产效率的同时, 使得独立工业生产终端不再成为一个相对安全的数据孤岛, 多元化的数据接入方式使得工业控制终端更加容易受到外界的攻击威胁[1]. 工业控制终端功能化的设计目标, 使其在设计时未能考虑安全防护的需要, 有限的计算存储资源也制约了安全防护措施的接入, 导致近年来以“震

① 基金项目: 国家科技重大专项 (2017ZX01030-201)
Foundation item: National Major Project of Science and Technology of China (2017ZX01030-201)
收稿时间: 2017-11-21; 修改时间: 2017-12-15; 采用时间: 2017-12-20; csa 在线出版时间: 2018-06-27

网”病毒为代表的工业网络安全事件时有发生[2]. 大量 0-day 漏洞的利用以及愈发丰富的变种攻击手段使得传统的基于漏洞库的安全防护策略暴露出更加多的局限性[3], 设计准确高效且能够应对未知类型攻击的安全防护策略正逐渐成为当前研究关注的焦点.

现如今, 工业控制网络行为有限、状态有限的特性已逐渐成为工控网络安全研究的切入点, 基于通信数据包的深度解析结果实现行为提取、链路检测的边界防护手段也已成为一种新的安全实现策略[4]. 在提取每一个数据包对应的操作将数据通信过程转换为行为序列进行建模分析时, 通常需要有标注的异常序列样本, 以建立起识别异常行为的序列标注问题模型[5], 但大量的有标注异常样本数据从实际生产环境中获得存在一定的难度. 单类支持向量机 (One Class Support Vector Machine, OCSVM) 能够仅使用单一类别样本实现二分类模型的建立, 为基于实际生产环境中的正常行为样本数据建立异常行为识别模型提供了有效的解决途径[6], 但为保留序列的上下文特性, 克服序列长短不一问题时, 所构造出的特征向量则会产生高维稀疏性问题.

针对实际生产环境中异常类别样本数据难以获得以及构建序列特征存在的高维稀疏性的问题. 本文采用了将语义向量模型[7]与单类支持向量机相结合的建模方式, 使用实际生产环境中的正常样本数据基于 OCSVM 实现对异常行为识别模型的构建. 通过语义向量模型将不同长度的控制行为序列转化为相同维度的特征向量, 保留序列中各控制行为间的上下文关系的同时满足常见分类模型的建模需求. 最后, 通过仿真各种常见的攻击方式构造出多种类型的异常行为序列作为测试数据集, 用以验证所构建的异常行为识别模型的准确性.

1 控制行为定义与异常识别定位

本文以电力 SCADA 系统这一工业控制网络的典型代表作为研究对象, 对其采用的 IEC 104 规约控制协议数据帧进行深度包检测结果定义控制操作. 使用时间窗划分法获得控制行为序列, 并以此作为分析识别的对象建立对异常行为序列识别模型, 实现出现异常控制行为时及时报警并定位异常位置.

1.1 协议解析与操作定义

IEC 104 规约是基于 TCP/IP 网络的远动设备与系统的通信传输标准, 该规约中报文帧格式包括定长帧与变长帧两种. 每一帧均是一个应用规约数据单元 (Applying Protocol Data Unit, APDU). 变长帧由应用规约控制信息 (Applying Protocol Control Information, APCI) 和应用服务数据单元 (Application Service Data Unit, ASDU) 组成[8]. APCI 的长度为 6 字节, 定义了报文传输启动/停止以及传输连接监视等控制信息, 控制报文的可靠传输. ASDU 由数据单元标识符和一个或多个信息对象所组成. 而定长帧则只包含 APCI 部分, 104 规约的帧格式如图 1 所示.

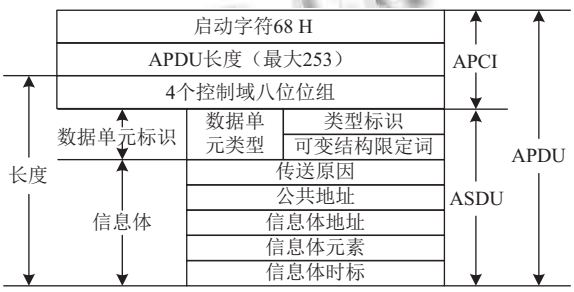


图 1 104 规约帧格式

根据工业控制网络行为有限、状态有限的特性, 对 104 规约数据包进行深度包检测, 可以提取其对应的控制操作集合$A=\{a_1,a_2,\cdots,a_i,\cdots,a_m\}$, 其中 m 为控制操作类型数. 对于以变长帧形式传输遥测、遥信、遥控、遥调等信息的 I 格式报文[9], 其 ASDU 部分中定义的 127 种数据包类型标识和 47 种传输原因, 则将每个类型的 I 格式帧对应的控制操作分别记为 a_1~a_{5969}.

对于不含 APCI 的定长帧格式传输的 S 格式报文和 U 格式报文, 根据图 2 中所示的控制域格式可知, 仅用于提供报文序号确认的 S 格式报文的控制操作被记为 a_{5970}, 用于完成 6 种传输规约控制的 U 格式报文的控制操作则被记为 a_{5971}~a_{5976}.

图 2 S 格式与 U 格式报文控制域

1.2 控制行为划分

在工业控制网络中, 每个工业控制主机到相应受控单元的生产业务可以抽象为一系列的控制操作序列. 当网络遭受外界的恶意挟持攻击时, 产生的业务异常通常体现在控制操作序列的异常. 采用时间窗划分将操作行为序列细分为描述控制行为的操作子序列 s_k, 并将其作为异常控制行为的识别对象, 既能够保留控制行为中包含原始操作, 也包含了行为中各具体操作的频率特性.

根据对通信数据包进行深度包检测提取到的源 IP 地址、目的 IP 地址、源端口、目的端口、控制行为类型的五元组信息: $<SrcIP, DestIP, SrcPort, DestPort, a_i>$, 将属于相同通信链路的数据包按照产生的先后顺序划分至同一行为序列中.

为避免由于时间窗划分对属于同一控制行为的连续控制操作的误分, 在以 15 s 为一个时间窗长度划分的基础上, 采用图 3 中所示的以 5 s 为一个增量单位的滑动时间窗口机制, 完成对控制操作子序列的提取, 确保对控制行为的准确描述.

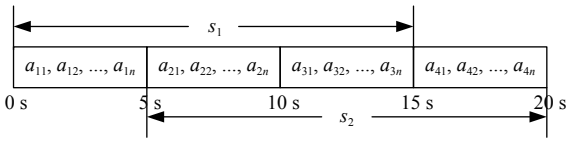


图 3 采用滑动窗口的序列划分

1.3 异常行为识别的建模过程

在工控网络的通信协议中, 控制行为的发出与响应过程有着较为严格约束, 一系列规范的控制操作组成了特定的控制行为. 结合各操作子序列 s_k 中各相邻控制操作间的上下文语义特性, 针对异常行为的序列识别建模分析需经过图 4 中所示的共计以下 6 个数据处理过程.

1) 对所抓取的数据包进行深度包检测, 提取包含通信链路与控制行为的五元组信息.

2) 根据通信链路进行数据混洗、合并, 按照时间窗划分出控制操作序列.

3) 将得到的控制行为序列进行语义向量建模, 获得序列的数值化、向量化表示.

4) 使用正常生产环境中控制行为序列的向量化样本数据, 采用单分类算法构建异常行为识别模型.

5) 将由 1)、2) 步提取的未知行为类型的操作序列, 经过语义向量模型转换为数值向量, 输入异常行为识别模型中获得识别结果.

6) 针对异常行为序列, 根据其对应的时间窗分片信息、通信链路 IP、通信链路端口号定位出现异常的时间、工作节点、业务应用, 寻找异常原因.

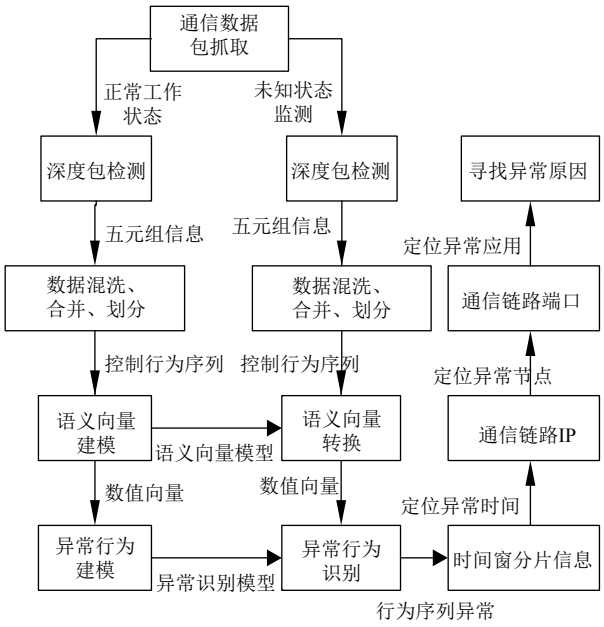


图 4 异常行为识别的建模分析过程

2 序列特征提取与行为识别建模

采用语义向量模型将各时间窗内不同长度的行为序列转换为统一维度的向量化表达满足异常序列识别的建模的需求. 基于单分类算法实现仅使用单类正样本完成异常行为识别的建模, 克服实际生产环境中异常行为序列难以获取的问题.

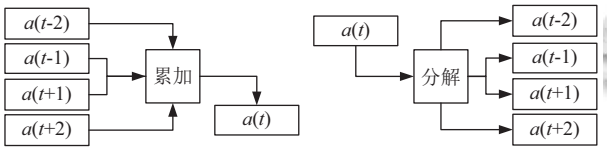
2.1 采用语义向量数值化行为序列

使用传统方式对操作序列进行向量化表达时, 通常统计序列中各控制操作或指定连续操作所出现的次数, 作为该序列的向量化表示[10]. 所获得的数值向量无法涵盖序列中相邻操作的上下文语义关系, 在子操作类型较多时所得的向量还会产生高维稀疏性问题.

为获得对控制行为序列准确的向量化表达, 结合相邻控制操作间的上下文语义特性, 使用 CBOW 模型和 Skip-gram 模型将各控制操作转换为包含具体操作含义的、在指定维度空间上的数值化向量化表达[11]. 并在此基础上, 构建 PV-DM 和 PV-DBOW 模型, 将行

为序列转换为包含语义特性的向量化表达.

CBOW 模型在给定序列中第 t 个操作前后 c 个操作的情况下预测第 t 个操作, 而 Skip-gram 模型则是给定第 t 个操作预测其前后 c 个操作[12]. 图 5 为 c=2 时两个模型的结构. 在 CBOW 模型中, 输入层为操作 w_t 前后 c 个操作对应的数值向量, 而投影层向量 X_w 为这 $2c$ 个向量的累加和, 输出层为包含 m 个叶子节点的 Huffman 树, 其中 m 为操作集合 A 中操作的总数, Huffman 树的编码则根据在整个训练集中各单词所出现的频率对应的权值进行构建. 同理, Skip-gram 模型的结构与 CBOW 模型的结构相似.



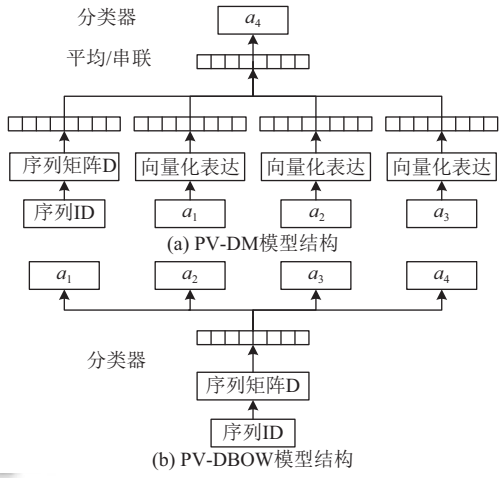
(a) CBOW 模型结构 (b) Skip-gram 模型结构

图 5 CBOW 与 Skip-gram 模型结构

两个模型的训练目标分别为对于每一个操作 a_t 使得 $P(a_t|a_{t-c},\cdots,a_{t-1},a_{t+1},,a_{t+c})$ 和$P(a_{t-c},\cdots,a_{t-1},a_t,a_{t+1},\cdots,a_{t+c}|a_t)$的概率值最大化. 使用随机梯度下降训练 CBOW 和 Skip-gram 两个神经网络模型的中间层参数 X_w 直至收敛, 最终获得各个操作最优的向量化表达.

在对长度各不相同的行为序列进行向量化表达时, 考虑序列内各操作具体含义的基础上, 还需要考虑序列中各操作的频率和操作之间的上下文关系. 在获得各子操作向量化表达的基础上, 采用相似的模型构建和优化手段, 构建图 6 中的 PV-DM 和 PV-DBOW 模型[13]. 模型中操作的向量化表达采用对 CBOW 模型和 Skip-gram 的训练优化结果. 按照相同的神经网络训练方式, 最终使得$P(s|a_1,a_2,\cdots,a_n)$和$P(a_n|a_1,a_2,\cdots,a_{n-1},s)$的概率值最大化, 即可得到各个行为序列最优的向量化表达.

使用所采集到的控制行为序列集合 S 作为训练数据集训练 CBOW 模型和 Skip-gram 模型获得每个控制行为 a_i 的向量化表达, 并在此基础上进一步训练 PV-DM 和 PV-DBOW 模型, 实现将行为序列转化为数值化向量表达. 将采集到控制行为序列集合 S 中的每一条行为序列 s_i 转化为 k 维特征向量 x_i, 即可获得用以构建异常识别模型的训练数据集 X, 其中 k 为语义向量模型中所指定的向量维度.



(a) PV-DM模型结构

(b) PV-DBOW模型结构

图 6 PV-DM 与 PV-DBOW 模型结构

2.2 基于 OCSVM 构建异常识别模型

在实际的生产环境中, 异常行为序列样本的获取存在一定难度. 基于已知的先验知识对异常控制行为进行仿真, 仅能获得有限的异常样本. 广泛应用于异常识别的传统支持向量机 (Support Vector Machine, SVM) 模型, 使用正负样本分布不均衡的数据集训练模型时同样会产生过拟合多数类样本的问题.

将 SVM 结合树形层次结构调整模型的训练过程, 对多数样本类数据进行聚类获得关键簇集[14], 用远离分类超平面簇的中心样本代替簇内样本, 能够消除多数类中非支持向量样本引起的样本不均衡性. 结合集成学习训练多个基分类器的策略, 可以进一步提升算法的泛化能力[15], 使模型对少样本类拥有同样准确性. 工业控制网络中, 异常行为没有明确的界定范围, 使用仿真异常样本训练出的异常识别模型对未知攻击类型的异常行为无法保证较低的漏报率. 因此, 采用单分类模型对正常样本在特征空间中分布的建模思想, 实现对未知样本是否属于正常状态的判断.

基于统计未知样本点附近正常样本点的数量[16], 衡量未知样本是否属于目标类别的单分类模型, 需要花费大量的存储计算开销计算与已知类别样本间的距离, 无法满足高响应速率的需求. 将行为序列转为向量化表达后, 各维特征的取值在样本空间中的分布未知, 无法适用于基于目标类型样本空间中密度分布建模的单分类模型[17].

支持向量描述方法 (Support Vector Data Describe,

SVDD) 采用在高维特征空间寻找包围所有目标类别样本点超球面的单分类模型思想, 采用与 SVM 相近的最小化样本点到超球面间隔的思想, 寻找描述边界超球面的支持向量. 在使用相同核函数的情况下与本文采用的 OCSVM 算法完全等价[18], 获得目标对偶问题和分类决策函数的进一步简化形式.

OCSVM 的主要思想是将单分类问题等价为特殊的二分类问题, 即使用全部属于同一类别的训练样本, 通过核函数将输入空间映射到高维空间, 寻找最优分类超平面, 将训练样本点尽可能与原点分开[19]. 使用高维空间中的分类超平面函数判断输入样本点是否属于已知类别, 其对应的二次优化问题如下:

$$\min(1/2)\|\omega\|^2+(1/vl)\sum_i^l\xi_i-\rho \tag{1}$$

$$\text{s.t. } \Phi(x_i)\omega\geq\rho-\xi_i\ \ \xi_i>0\ \ i=1\cdots l \tag{2}$$

训练样本 $x_1,x_2,\cdots,x_l\in X$, l 是训练样本总数, Φ: $X\to H$ 是原始特征空间到高维空间的映射, ω 和 ρ 分别为特征空间中所需超平面的法向量和补偿. $v\in(0,1]$是用以控制数据集中异常样本比例上界与支持向量比例下界的权衡参数, 松弛变量 ξ_i 为允许训练样本被错误分类的程度.

最终获得代表分类超平面的决策函数为:

$$f(x)=\text{sgn}(\Phi(x)\omega-\rho) \tag{3}$$

引入拉格朗日函数将上述二次规划问题转换为:

$$\begin{aligned}L(\omega,\xi,\rho,\alpha,\beta)=&(1/2)\|\omega\|^2+(1/vl)\textstyle\sum_i^l\xi_i-\rho\\&-\textstyle\sum_i^l\alpha_i(\omega\cdot x_i-\rho+\xi_i)-\sum_i^l\beta_i\xi_i\end{aligned} \tag{4}$$

对 ω,ρ,ξ_i 分别求偏导可得:

$$\begin{aligned}&\omega=\textstyle\sum_i^l\alpha_i\Phi(x_i)\\&\alpha_i=1/vl-\beta_i\leq1/vl,\textstyle\sum_i^l\alpha_i=1\end{aligned} \tag{5}$$

其中, α_i,β_i 分别为拉格朗日乘子. 并引入高斯核函数:

$$K(x_i,x_j)\leq\Phi(x_i),\Phi(x_j)\geq\exp(-g\|x_i-x_j\|^2) \tag{6}$$

其中, g 为高斯核函数参数, 将公式 (5)(6) 代入式 (4) 中得到其对偶问题为:

$$\min_\alpha(1/2)\sum_i^l\sum_j^l\alpha_i\alpha_jK(x_i,x_j) \tag{7}$$

$$\text{s.t. } 0\leq\alpha_i\leq1/vl,\ \ i=1,\cdots,l,\ \ \sum_i^l\alpha_i=1 \tag{8}$$

选取任一满足 $0\leq\alpha^*\leq1/vl$ 的 α^*, 计算出偏移量:

$$\rho=\sum_i^l\alpha_i^*K(x_i,x_j) \tag{9}$$

满足 $0\leq\alpha^*\leq1/vl$ 的向量叫支持向量, 最终求得决策函数如公式 (10) 所示, 其中 N_{SV} 为支持向量个数.

$$f(x)=\text{sgn}(\sum_i^{N_{SV}}\alpha_i^*K(x_i,x_j)-\rho) \tag{10}$$

基于 OCSVM 实现对异常行为识别模型的建立过程中, 将从正常生产状态下通过时间窗划分抽取到的多条行为序列 s_i 作为训练数据集 S, 采用所构建的文本模型将其转换为指定 k 维的特征向量 x_i, 基于训练样本集 X 得到的 OCSVM 模型即可实现对所输入的特征向量是否属于正常类型的识别.

对于未知类型的行为序列 s', 将其经过语义模型转为向量化表示后, 将所得的特征向量 x'代入所训练模型的决策函数 $f(x)$ 中, 输出该特征向量所属的类别, 实现对异常行为序列的识别.

3 仿真实验分析

3.1 实验环境与评估指标

本文采用的实验环境是由一台采用 IEC 104 规约进行通信的控制主机仿真器和一台受控单元仿真器组成的业务控制系统, 并向网络中接入流量传感器模块对数据包中的操作行为进行解析, 整理汇总出各条通信链路的行为序列.

通过调整仿真器的工作模式和所仿真终端的类型用以模拟包括遥信、遥控、遥测、遥调等多种正常工作状态下的控制行为, 收集共计 16 000 条的正常行为序列. 在劫持控制终端后, 针对受控单元的攻击主要包括随机操作、篡改行为、重复指令、颠倒业务、未知指令等多种方式. 因此, 在所采集的正常行为序列的基础上通过随机构造、复制、裁剪、易序、伪造等手段, 仿真以上 5 种攻击类型的行为序列各 200 条, 获得共计 1000 条异常行为序列.

为验证语义向量模型结合 OCSVM 算法对异常行为序列识别的准确性, 使用所获得的 15 000 条正常行为序列作为训练数据集. 测试数据集则由正常行为序列和异常行为序列各 1000 条构成, 并采用以下两个指标评估异常行为识别模型在测试数据集上的准确性:

准确率: Precision=TP/(TP+FP)

召回率: Recall=TP/(TP+FN)

其中 TP (True Positive) 表示识别为异常序列中识别正确的数量, FP (False Positive) 表示识别为异常序列中识别错误的数量, FN (False Negative) 则表示实际为异

常行为序列但识别为正常的数量.

3.2 识别准确性分析

基于规则的异常行为识别策略通常采用已知的非法行为构建用以进行异常模式匹配的操作子序列集合. 结合实际业务中对非法行为操作的定义, 构造出由 423 条非法子序列构成的模式匹配集合, 作为与本文采用的异常行为识别算法的对照.

为验证语义向量模型对异常行为识别的准确性提升, 将基于操作和操作组合频率统计的传统行为序列的向量化方式作为对比. 同时, 使用 LDA 话题模型对传统方式提取的特征向量进行降维[20], 将序列中各行为加权频率转化为在各个抽象“话题”上的分布权重, 并采用 OCSVM 完成异常行为识别建模.

其中, 语义向量模型中行为序列特征向量对应的目标维数 K=50, LDA 话题模型的目标维数同样为 50, OCSVM 算法模型参数分别取各自在训练数据集上的最优参数, 并按照所使用的三种向量化方式分别对测试数据集中的数据进行向量化操作.

将测试集正常样本五等分, 与 3.1 节中 5 种攻击类型的异常行为序列分别构成 5 个测试子集. 分别由三种特征构建方法所得特征向量训练的 OCSVM 模型与传统异常模式匹配策略在各测试子集上的性能评估指标分别如表 1 所示.

表 1 不同特征构造方式下的异常识别准确性 (单位: %)

测试集合	异常子序列模式匹配		传统方式 & OCSVM		传统方式 & LDA & OCSVM		语义向量 & OCSVM	
	准确率	召回率	准确率	召回率	准确率	召回率	准确率	召回率
随机操作测试子集	85.8	55.6	78.1	79.2	88.6	88.2	94.1	94.5
篡改行为测试子集	86.5	57.5	80.9	78.5	89.5	84.1	96.9	95.3
重复指令测试子集	89.3	62.3	72.3	72.7	84.4	85.5	91.3	92.7
颠倒业务测试子集	88.2	59.6	70.2	71.1	85.7	87.6	94.7	93.4
未知指令测试子集	87.4	51.2	76.5	74.3	86.7	86.8	90.4	92.9
总测试集	88.1	57.2	76.8	75.2	87.3	86.4	93.2	93.8

传统向量化方式所构造的特征存在高维稀疏性, 所得的向量中大部分特征维度的值为 0, 将未进行降维的特征向量用于训练所得到的异常识别模型的准确性较低. 相较于传统向量化方式结合 LDA 降维算法, 语义向量模型所构造的特征向量在考虑时间序列中控制行为频率特性的同时, 保留了控制行为之间的上下文关系, 使模型准确性进一步提高.

尽管基于规则的异常模式匹配策略对所识别出的异常行为有不错的准确率, 但其根据先验知识的匹配检测思路使其在测试数据集上的召回率较低, 即存在大量漏报的情况, 因此无法满足实际应用场景中对未知异常行为准确识别的需要.

从随机操作、篡改行为、重复指令三个测试子集中各取 100 条异常样本数据加入训练集, 采用树形层次结构与 AdaBoost 对传统 SVM 模型进行优化作为与 OCSVM 模型的对比, 训练数据的向量化方式均采用语义向量模型. 由表 2 中结果可知传统 SVM 模型的准确性受正负样本不均衡性的影响较大. 采用树形层次结构与 AdaBoost 优化后的模型有效克服了正负样本不均衡对模型训练的影响, 但在颠倒业务和未知指令两个测试集上的准确率较低并存在较为明显的漏报, 无法满足对未知攻击类型异常行为的识别需要.

表 2 各类支持向量机异常识别的准确性 (单位: %)

测试集合	随机操作测试子集		篡改行为测试子集		重复指令测试子集		颠倒业务测试子集		未知指令测试子集	
	准确率	召回率	准确率	召回率	准确率	召回率	准确率	召回率	准确率	召回率
传统 SVM	72.5	73.6	68.3	69.2	72.4	71.7	74.1	73.8	68.5	69.2
树形层次结构+SVM	92.3	93.7	92.6	93.3	92.5	93.4	87.5	78.2	88.6	79.1
Adaboost+SVM	92.9	90.2	94.5	92.2	89.9	91.1	84.8	77.6	82.5	75.3
OCSVM	94.1	94.5	96.9	95.3	91.3	92.7	94.7	93.4	90.4	92.9

3.3 算法计算开销比对

模型构建与识别的计算开销是模型能否满足实际应用需要另一衡量标准. 分别采用特征提取过程中向量化建模与行为序列向量化的耗时、OCSVM 建模部分的迭代轮次与单位轮次迭代耗时及模型识别响应耗时 5 个指标对 3.2 节中三类模型的时间开销进行对比. 分别对各模型进行 5 次相同的建模计算与响应过程, 对各阶段具体的耗时取平均其结果如表 3 所示.

表 3 各模型不同阶段的计算开销

计算开销 统计指标	传统方式 &OCSVM	传统方式 & LDA & OCSVM	语义向量 & OCSVM
向量化建模耗时 (s)	0.847	0.862	3.416
行为序列向量化耗时 (s)	0.269	2.374	0.458
OCSVM 建模迭代轮次	3679	1922	1763
单位迭代耗时 (100 轮)(s)	1.206	0.712	0.745
识别响应耗时 (1000 条)(s)	1.374	0.937	0.952

由实验结果可知, 尽管传统方式在向量化建模和行为序列向量化过程中的耗时均较低, 但其构建特征向量的高维稀疏性使模型训练过程中的单位迭代耗时和迭代收敛轮次均大于其他方式. 使用语义向量模型所得的特征向量训练时需要更少的迭代轮次使模型趋于收敛, 尽管特征提取时产生了一定的时间开销, 但仅占总开销的一小部分. 同时, 所构建的 OCSVM 模型对异常行为序列的响应时间符合实际应用的需要.

4 结论与展望

本文以电力 SCADA 系统中常用的 IEC 104 规约通信协议为例, 通过对数据包内容进行深度解析, 根据不同数据包所对应的控制操作类型, 将生产业务过程抽象为控制行为序列进行建模实现对异常行为序列的识别. 根据工控网络协议的语义特性和数据包之间的上下文关系, 采用语义向量模型将各时间窗内长度不同的行为序列转换为相同维度的特征向量. 基于 OCSVM 算法实现了在仅使用正常样本的条件下对异常行为的识别实现准确建模. 使用多种类型的行为序列验证了所构造的模型对异常序列、未知序列的识别具备较高的准确性. 下一步将对单分类模型在异常行为识别的可靠性和准确性上进行进一步的优化提升.

参考文献

1 胡毅, 于东, 刘明烈. 工业控制网络的研究现状及发展趋势. 计算机科学, 2010, 37(1): 23–27, 46.

2 杨安, 孙利民, 王小山, 等. 工业控制系统入侵检测技术综述. 计算机研究与发展, 2016, 53(9): 2039–2054. [doi: 10.7544/issn1000-1239.2016.20150465]

3 赖英旭, 刘增辉, 蔡晓田, 等. 工业控制系统入侵检测研究综述. 通信学报, 2017, 38(2): 143–156. [doi: 10.11959/j.issn.1000-436x.2017036]

4 张盛山, 尚文利, 万明, 等. 基于区域/边界规则的 Modbus TCP 通讯安全防御模型. 计算机工程与设计, 2014, 35(11): 3701–3707. [doi: 10.3969/j.issn.1000-7024.2014.11.001]

5 刘帅. 面向网络数据流的多层面异常行为分析检测技术研究[硕士学位论文]. 郑州: 解放军信息工程大学, 2015.

6 尚文利, 张盛山, 万明, 等. 基于 PSO-SVM 的 Modbus TCP 通讯的异常检测方法. 电子学报, 2014, 42(11): 2314–2320. [doi: 10.3969/j.issn.0372-2112.2014.11.029]

7 Le Q, Mikolov T. Distributed representations of sentences and documents. Proceedings of the 31st International Conference on International Conference on Machine Learning. Beijing, China. 2014. II-1188–II-1196.

8 韩滨旭, 沈玉玲, 庞海亮. 组态软件中的 IEC60870-5-104 规约实现. 制造业自动化, 2015, 37(12): 114–118.

9 李宣义, 梁宾, 李均强, 等. 基于变电站综合自动化调试试验系统的 104 规约一致性测试研究. 河北电力技术, 2016, 35(6): 46–48, 62.

10 尚文利, 李琳, 万明, 等. 基于优化单类支持向量机的工业控制系统入侵检测算法. 信息与控制, 2015, 44(6): 678–684.

11 黄仁, 张卫. 基于 word2vec 的互联网商品评论情感倾向研究. 计算机科学, 2016, 43(6A): 387–389. [doi: 10.11896/j.issn.1002-137X.2016.6A.092]

12 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. arXiv:1301.3781, 2013. 1–12.

13 唐明, 朱磊, 邹显春. 基于 Word2Vec 的一种文档向量表示. 计算机科学, 2016, 43(6): 214–217, 269. [doi: 10.11896/j.issn.1002-137X.2016.06.043]

14 邓曦辉, 赵丽. 树形层次结构的非平衡 SVM 分类方法. 计算机工程与设计, 2017, 38(8): 2269–2275.

15 杨云, 卢美静. 基于集成支持向量机的葡萄酒品质分类方法. 计算机工程与设计, 2017, 38(9): 2541–2545.

16 薛文, 谢永红, 马延辉, 等. 基于集成金字塔模型的单分类方法. 计算机科学, 2011, 38(6): 191–194.

17 孙强, 魏伟, 侯培鑫, 等. 基于区间数单簇聚类-单分类器的异常检测. 计算机科学, 2017, 44(6): 189–198, 205.

18 王洪波. 单分类支持向量机的学习方法研究[博士学位论文]. 杭州: 浙江大学, 2012.

19 李琳, 尚文利, 姚俊, 等. 单类支持向量机在工业控制系统入侵检测中的应用研究综述. 计算机应用研究, 2016, 33(1): 7–11.

20 王振振, 何明, 杜永萍. 基于 LDA 主题模型的文本相似度计算. 计算机科学, 2013, 40(12): 229–232. [doi: 10.3969/j.issn.1002-137X.2013.12.049]