doi:10.3772/j.issn.1002-0470.2021.02.001

基于监督学习的规则触发执行预测方法研究①

黄晓辉②*** 崔 莉③** 黄 希**

(*中国科学院大学 北京 100049)

(**中国科学院计算技术研究所 北京 100190)

摘 要 针对物联网(IoT)规则引擎在实际应用中,由于外部、内部用户以及系统本身的不安全因素引起的控制及执行安全问题,提出了一种基于监督学习理论的规则触发执行设备控制预测方法(EasiPRED)。首先,制定了物联网规则引擎的规则统一描述标准,便于规则的设置和管理;然后,基于主成分分析(PCA)方法提出了一种特征提取算法,分别对规则样本集进行降维以及规则特征提取,不仅保留了原有的特征信息,同时也降低了计算复杂度;最后,基于监督学习方法对规则样本集进行训练学习,预测规则在触发执行后的控制类别,并根据预测结果与实际规则的输出做一致性验证,以此判断控制行为的安全性。实验结果表明,EasiPRED 能够有效地识别出不安全的控制规则,提高系统的整体安全性能,避免发生控制及执行安全事故。

关键词 物联网(IoT)规则引擎; 不安全因素; 控制安全; 监督学习; 控制类别

0 引 言

物联网(Internet of Things, IoT)作为一个新兴的信息技术领域,自提出以来就已成为许多国家发展的战略[1]。规则引擎是一种嵌入在应用程序中的组件,能够方便、灵活地将业务规则从实现逻辑中分离出来,并使用预定义的语义模块编写业务决策[2],能够方便、灵活地将规则内容从实现逻辑中分离出来,因此在工业和学术界都得到了广泛的关注[3]。随着社会对物联网应用需求的日益增长,如智能医疗、环境监测与保护、智能家居等,具备规则引擎模块的物联网平台,能够实现物联网设备数据的不同处理需求,如实时告警、联动控制等操作。

伴随着物联网规则引擎在各行各业的进一步推广应用,由外部用户入侵、开发者的错误编码、内部用户的误操作以及系统本身故障等引起的不安全控制因素(unsafe control factors, UCF)影响规则的正确触发执行后,会因为设备的错误控制导致系统发生控制安全事故,甚至直接影响到用户的生命和财产安全。针对上述问题,许多研究从用户和系统的不同角度出发,通过采用身份识别、访问控制、行为预测以及故障检测等方法来解决这类安全问题。

在身份识别方法中,文献[4]提出了双因素身份认证协议,提供强大的身份验证和会话密钥建立过程,该协议抵制多用户相同身份登录的威胁,身份验证过程不需要公钥操作,通过加密散列函数达到较高的认证效率。文献[5]提出了一种创新型的应用 PUF (physical unclonable function)用于改善熵,并将基于 PUF 的熵泵应用于增强嵌入式物联网设备的密码,实证了人工生成密码的熵值总体上提高了48%。文献[6]提出了基于身份的密钥协商算法,该算法具有隐式的双向身份验证的功能,并以该算法为基础,设计了物联网中订阅分发架构下的双

① 国家自然科学基金(61672498)和国家重点研发计划(2016YFC0302300)资助项目。

② 男,1987 年生,博士生;研究方向:物联网,无线传感器网络;E-mail:huangxiaohui@ict.ac.cn

③ 通信作者,E-mail: lcui@ict.ac.cn

(收稿日期:2020-02-24)

向身份验证协议,利用信任链中信任关系的可传递性证明了设备可以间接完成双向身份验证。文献[7]提出了基于多密钥的相互认证机制,该方法中的服务器和物联网设备之间共享的秘密被称为安全库,它是一组大小相同的密钥,安全库的初始内容在服务器和物联网设备之间共享,安全库的内容在每次成功通信会话后都会更改。

文献[8]侧重于访问控制方法的研究,利用椭圆曲线密码技术(elliptic curve cryptography, ECC)和抗碰撞的单向密码散列函数,提出了一种新的基于证书的轻量级访问控制和密钥协议,并在 ROR(real-or-random)模型下,使用安全协议分析工具 AVISPA(automated validation of Internet security protocols and applications)进行了详细的分析,表明该协议在需要一个安全的访问控制机制时能够防止物联网设备受到各种已知的攻击,在安全性、功能性、通信和计算成本方面有更好的权衡性能。另外,文献[9]提出了基于角色层次访问控制的代理实现方案,适用于物联网中基于角色的访问控制,该方案具有计算效率高、所需内存少、实现成本低等优点,适用于具有安全层次密钥管理的物联网设备。

在行为预测方法上,研究工作者通过预测用户的下一步行为,来判断行为是否符合规律。文献[10]提出了一种改进的隐马尔可夫模型(improved hidden Markov model, IHMM)来支持残疾人的个性化行为预测。该模型以不同的日温度区间值被表征和识别为隐含变量,通过打破隐马尔可夫模型的时不变假设,将时间信息作为状态的位置,建立时间状态转移矩阵代替固定状态转移矩阵来证明状态转移的概率,并指导用户的活动。对所提工作的评估表明,与传统的隐马尔可夫模型相比,隐马尔可夫模型至少提高了 10% 的预测精度。文献[11]提出了改进差分整合移动平均自回归模型(auto regressive integrated moving average , ARIMA)预测用户行为的方法,以热水器开机时间为例,通过对热水器连续 12 周开机时间的原始数据进行分析,建立了一种改进的 ARIMA 模型来预测用户在第 13 周开机的时间。将改进后的 ARIMA 模型与 ARIMA 模型进行比较,结果表明改进后的 ARIMA 模型具有较好的预测性能,能满足短期准确预测的需求。

在故障检测方法上,文献[12]提出了基于趋势自适应多尺度主成分分析(trend-adaptive multi-scale principal component analysis, TAMS-PCA)的数据故障检测模型,该模型继承了离散小波变换(discrete wavelet transform, DWT)在获取时频信息方面的优势和主成分分析(principal component analysis, PCA)在提取传感器数据间相关性方面的优势,并在真实数据集上的实验结果表明了该模型在数据故障检测中的有效性。文献[13]为了感知不同硬件组件的故障,采用了基于软件开销的最小故障检测方法,通过对故障节点准确及时的检测,提高了无线传感器网络(wireless sensor network, WSN)的可靠性,提出了一种基于聚类的故障检测算法,提高了系统的可扩展性,减少了网络检测的负载。

现有的研究工作中,身份识别、访问控制和行为预测的工作主要考虑的是用户方面,通过识别用户的身份、访问权限以及预测用户的下一步行为来减少不安全控制因素的发生,而故障检测方法则考虑的是系统本身设备的不安全因素,即设备故障引起的控制安全问题。上述方法从单方面角度考虑,都能够通过降低用户和系统本身的不安全控制因素来避免控制安全问题,而本文则是以规则为出发点,综合考虑了用户和系统所存在的不安全控制因素。不论是用户还是系统本身的原因造成的控制安全问题,在物联网规则引擎中,最后都是体现在规则的触发执行上。

综上所述,在给定的规则触发条件下执行的控制是否符合规则触发执行规律,这也是保障物联网规则引擎控制安全的一种新思路,因此,本文提出了一种基于监督学习的规则触发执行预测方法(EasiPRED)。本文的主要贡献包括:(1)制定了物联网规则引擎的规则统一描述格式,便于规则的设置和管理。(2)基于主成分分析和规则特征属性,提出了一种规则特征提取算法,在保留原规则特征信息的前提下,降低了规则特征向量的复杂度。(3)提出了一种规则触发执行预测算法,通过挖掘规则触发信息与执行控制类别之间的关联性,预测规则触发执行后的设备控制类别是否与实际执行控制结果

一致,据此判断执行规则控制的安全性,并通过实验验证了 EasiPRED 在物联网规则引擎实际应用中降低系统发生控制安全事故的有效性与可行性。

1 物联网规则引擎概述

1.1 物联网规则引擎结构

物联网平台引入规则引擎后,使得用户能够对接入平台的物联网设备设定相应的规则,在条件满足所设定的规则后,平台会触发相应的控制动作来满足用户需求,快速实现数据转发和设备联动控制[14]。规则引擎的核心结构如图 1 所示,主要包括控制安全模块、规则库(rule base, RB)、规则管理(rule management, RM)、推理引擎(inference engine, IE)、议程(agenda)、工作内存(working memory, WM)、规则编辑环境(rules authoring environment, RAE)、外部组件(external component, EC)等部分。控制安全模块用于预测规则触发执行后的设备控制类别,并与实际规则在触发执行后的设备控制类别进行一致性验证;规则库中存储着用户编辑的所有规则,采用 if-then 结构,if 部分描述规则的触发条件,then 部分为规则的执行部分,包括了设备信息和执行动作;RM 用于管理规则库中的规则;议程模块对 RM 取出的可执行规则进行处理,根据特定的排序方法确定执行顺序;推理引擎是根据规则 then 执行部分的具体描述来完成对对应执行设备的控制动作;工作内存主要包含用于规则匹配的事件;规则编辑环境,是用户编辑规则的主要环境;外部组件,可添加其他功能组件。用户在规则编辑环境中定义规则后存入规则库中,规则管理模块从规则库中取出规则,然后由规则执行模块和推理模块,将触发数据与规则的触发条件相比对,满足条件则执行相应操作[15]。

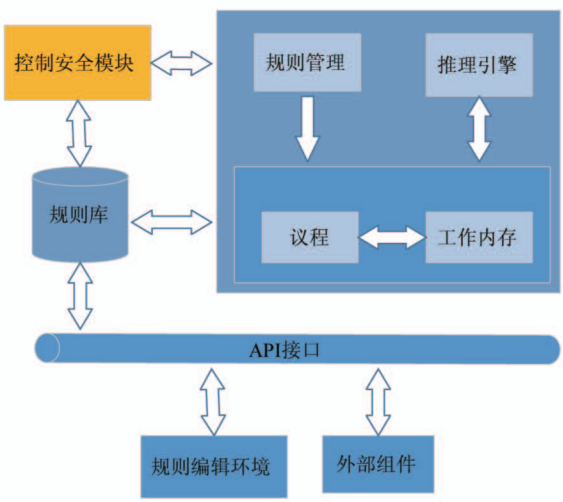

图 1 物联网规则引擎核心结构

1.2 规则统一描述

物联网中的各种感知信息组成了海量的泛在资源,存在着描述难、检索难、运营难等困扰统一服务的问题。如果物联网规则引擎中的规则无法做到统一描述,规则的执行单元就会因为要兼容千变万化的物联网资源而变得庞大和复杂。很多物联网设备都具备自动化管理能力,例如家用照明灯可以根据室内的光照强度实现自动开关,空调可根据自身的温度和湿度传感器数据自动调控室内温度和湿度。设备的自动化控制规则可以直接烧录到设备中,但是一旦规则有变化就必须给每个设备重新烧录程序,使得规则难以管理,所以物联网设备的控制规则由云端统一设置和管理,能够更好地维护设备的控制规则。但是,随着规则量的增多,使得规则的设置和管理变得越来越复杂,为了更好地描述和管理日益增多的规则,就需要制定规则统一描述格式。根据文献[16]的研究工作,本文从规则属性、运行状态、触发条件以及执行部分来定义规则。规则中的各个参数描述如表 1 所示。

由文献[15]和表 1 可知,规则属性包括 *Rule _ ID*、*Rule _ Typ*、*Rule _ Pro*,规则状态 *Rule _ Sta* = 0, 1,表示规则是否被激活,触发条件包括 *Trigger _ Typ*、*Trigger _ ID*、*Trigger _ Value*、*Value _ Au*,执行部分包括 *Actuator _ ID*, *Actuator _ Value*。下面具体介绍一个规则来说明规则的描述。例如,一个规则为“当室外 PM2.5 指数大于 100 μg/m^3 时,关闭窗户”,设定规则 *ID* 为 10,执行优先级为 5,PM2.5 传感器 *ID* 为 15,窗户 *ID* 为 10,那么执行部分之前的参数可表示为{10, 1, 5, 0, 1, 15,100, 1},执行部分的参数可表示为{10, 1},最后通过规则的统一描述定义,规则可表示为 *Rule* = {10, 1, 5, 0, 1, 15, 100,1,10,1}。另外,当出现需要2个或多个传感

表 1 规则表示参数描述[16]

参数	描述	取值
Rule _ ID	规则唯一标识号	$0 \sim 2^{15}$
Rule _ Typ	规则类型	$0 \sim 2^{10}$
Rule _ Pro	执行优先级	$0 \sim 2^{8}$
Rule _ Sta	规则的触发状态:触发、等待	0, 1
Trigger _ Typ	触发类型:时间、时间	0, 1
Trigger _ ID	触发传感器唯一标识号	$0 \sim 2^{8}$
Trigger _ Value	触发的数据信息	$0 \sim 2^{16}$
Value _ Au	触发条件的数据值与设定的触发数据值的关系:小于、等于或大于	$0 \sim 2^{2}$
Actuator _ ID	执行设备的唯一标识号	$0 \sim 2^{10}$
Actuator _ Value	执行设备的控制参数:打开或关闭	0, 1

器同时触发规则并对同一设备执行控制时,首先通过 *Rule _ Typ* 判断规则是否为多条件触发规则,然后需要同时满足 *Rule _ Typ*、*Rule _ Pro*、*Actuator _ ID* 以及 *Actuator _ Value* 值一致,并且多个传感器数据需同时达到规则的 if 触发条件,例如,规则“当室内温度小于 20 ℃并且室外 PM2.5 指数大于100 μg/m^3 时,关闭窗户”,先分解成规则 1“当室内温度小于 20 ℃,关闭窗户”和规则 2“当室外 PM2.5 指数大于 100 μg/m^3 时,关闭窗户”,两个规则要同时满足触发条件,才能够对窗户执行关闭操作。

2 EasiPRED 设计与实现

本文所提出的基于监督学习的规则触发执行预测方法 EasiPRED 包括规则特征提取和 EasiNB 规则执行预测算法。下面分别介绍这两种算法。

2.1 规则特征提取

从 1.2 节的规则统一描述格式分析,前 5 个规则表示参数主要是对规则固定属性的描述,后 5 个参数是对触发的条件和执行内容的描述,其中 *Trigger _ ID*、*Trigger _ Value*、*Value _ Au* 是触发的条件,*Actuator _ ID*、*Actuator _ Value* 是规则执行内容部分。EasiPRED 的特征向量从规则的参数中选取,而多数规则参数都有特定的属性意义,所以选择哪些参数属性作为特征向量会直接关系到预测结果的准确率。参数 *Actuator _ ID*、*Actuator _ Value* 分别对应规则执行时对应的设备号和控制参数,可作为最后要预测的规则触发类别。为了减少算法复杂度,将这 2 个参数组合成的不同控制类别进行统一分类,比如{10, 1}和{10, 0}视为控制类别 1 和控制类别 2。

由表 1 可知,前 8 个参数都可作为输入的特征向量,但是由于各个参数在规则中起到的实际作用,也可以说贡献度是不一致的,可以通过减少特征向量的维度来减少算法的计算复杂度。本文提出了基于主成分分析的规则特征参数选择算法 EasiPCA,主成分分析 PCA 是一种常见的数据降维方法,主要目的是在原有数据组成参数“信息”损失较小的前提下,将高维的数据转换成低维数据,从而减小计算开销。EasiPCA 的设计思想如下。

设 x_1, x_1, …, x_n 为 n 个原始特征,设新特征 δ_i, $i=1, 2, \cdots, n$ 为原始特征的线性组合:

$$\delta_i = \sum_{j=1}^{n} a_{ij} x_j = \boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{x} \tag{1}$$

为了统一 δ_i 的尺度,可以设系数的模为 1,即:

$$\boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{a}_i = 1 \tag{2}$$

因此,可以将式(1)表示为矩阵形式:

$$\boldsymbol{\delta} = \boldsymbol{A}^{\mathrm{T}} \boldsymbol{x} \tag{3}$$

其中,$\boldsymbol{\delta}$ 是由新特征 δ_i 组成的向量,$\boldsymbol{A}$ 是特征变换矩阵,此处要解的是最优正交变换 $\boldsymbol{A}$,使得新特征 δ_i 的方差达到最大。正交变换保证了新的规则特征之间不相关,而新的规则特征方差越大,规则样本在对应的特征上的差异就越大,这一特征就越重要,进而最大程度保留了原规则特征信息。通过 EasiPCA 将原规则样本组成的特征向量降维,得到的新规则特征向量,可作为下一节讨论的 EasiNB 算法的输入量。

2.2 EasiNB 规则触发执行预测算法设计

深度学习近几年已经成为了机器学习中较为热门的子领域,并在语音识别[17]、图像识别等领域表现出了较高的性能[18]。然而,在小数据集以及需要根据实际应用需求来调整参数或更改模型设计的情况下,端到端深度学习方法[19]存在局限性,仍需要采用经典机器学习方法,因此,本文不适合采用深度学习方法来实现规则触发执行的预测。

在概率论与统计学中,贝叶斯算法基于与一个事件相关的条件先验知识,确定了该事件发生的概率,算法的核心设计适合本文的研究内容,因此,本

文基于朴素贝叶斯监督学习理论,提出了规则触发执行预测算法 EasiNB。

对规则的描述参数以及提取特征向量本文在 1.2 和 2.1 节中已经详细介绍过了,可以由 *Char _ 1*、*Char _ 2*、*Char _ 3* 组成的特征向量作为输入量,规则控制参数 *Actuator _ ID*、*Actuator _ Value* 组成的控制类型作为结果输出,下面介绍 EasiNB 的设计原理。

定义 1 $\boldsymbol{D}=(x_1^{(m)}, x_2^{(m)}, \cdots, x_n^{(m)}, y_n)$ 为规则样本,其中 $x_1^{(m)}, x_2^{(m)}, \cdots, x_n^{(m)}$ 为规则样本的 n 个属性特征,m 为规则样本数,$y_n=c_k$, c_1, c_2, $\cdots$, c_k, y_n 表示规则执行的设备控制类别。

根据上述定义,设定以下规则样本:

$$\begin{aligned}&(x_1^{(1)}, x_2^{(1)}, \cdots, x_n^{(1)}, y_1)\\&(x_1^{(2)}, x_2^{(2)}, \cdots, x_n^{(2)}, y_2)\\&\cdots\\&(x_1^{(m)}, x_2^{(m)}, \cdots, x_n^{(m)}, y_n)\end{aligned}\tag{4}$$

即有 m 个规则样本,每个规则样本有 n 个特征,特征输出有 k 个类别,定义为 c_1, c_2, $\cdots$, c_k。

由概率分布计算原理,可以得到规则的概率联合分布 $P(X, Y)$,如下所示:

$$\begin{aligned}&P(X, Y=C_k)=P(Y=C_k)P(X=x\mid Y=C_k)\\&=P(Y=C_k)P(X_1=x_1, X_2=x_2, \cdots, X_n=x_n\mid Y=C_k)\end{aligned}\tag{5}$$

式中,$P(X_1=x_1, X_2=x_2, \cdots, X_n=x_n\mid Y=C_k)$ 是一个复杂的 n 个维度的条件分布概率,计算复杂度较高。为了能够采用朴素贝叶斯模型来简化式(5)的计算过程,首先要确保样本之间是相互独立的,才不会影响最后的预测结果。本文通过计算每个规则样本与其他规则的皮尔逊相关系数(Pearson correlation coefficient, PCC),设定系数的阈值 d,所有规则样本超过该阈值的就剔除出样本空间,在误差允许的范围内,最后剩下来的规则样本可以视为相互独立的。式(5)可根据朴素贝叶斯定理得到简化的联合概率分布:

$$\begin{aligned}&P(X_1=x_1, X_2=x_2, \cdots, X_n=x_n\mid Y=C_k)\\&=\prod_{j=1}^{n}P(X_j=x_j\mid Y=C_k)\end{aligned}\tag{6}$$

EasiNB 的输出结果是规则触发执行属于哪种设备控制类型,设 C_{result} 是 $P(Y=C_k\mid X=X^{(test)})$ 最大化的规则执行的设备控制类别概率,表达式如下。

$$\begin{aligned}C_{result}&=\underbrace{argmax}_{C_k}P(Y=C_k\mid X=X^{(test)})\\&=\underbrace{argmax}_{C_k}P(X=X^{(test)}\mid Y=C_k)P(Y=C_k)\\&\quad/P(X=X^{(test)})\end{aligned}\tag{7}$$

由于式(7)的分母都是 $P(X=X^{(test)})$,并根据式(6)的结论,式(7)可以简化为

$$C_{result}=\underbrace{argmax}_{C_k}P(Y=C_k)\prod_{j=1}^{n}P(X_j=x_j\mid Y=C_k)\tag{8}$$

所以只要计算出 $P(Y=C_k)$ 和 $P(X_j=X_j^{(test)}\mid Y=C_k)(j=1, 2, \cdots, n)$,就能够得到规则样本的输出属于哪种规则执行设备控制类型的最大概率值。下面讨论怎么通过训练集计算这两个概率。

设 H 为规则样本输出控制类别为 C_k 的总数,对于 $P(Y=C_k)$,可通过极大似然估计得到 C_k 出现的概率,即:

$$P(Y=C_k)=H/m\tag{9}$$

对于 $P(X_j=X_j^{(test)}\mid Y=C_k)(j=1, 2, \cdots, n)$,这个主要取决于规则的先验条件。

根据规则的定义,可知表示规则的 X_j 是离散的值,假设 X_j 符合多项式分布,可以得到 $P(X_j=X_j^{(test)}\mid Y=C_k)$ 是规则执行设备控制类别为 C_k 时,$X_j^{(test)}$ 出现的频率,即:

$$P(X_j=X_j^{(test)}\mid Y=C_k)=m_{kj}^{test}/m_k\tag{10}$$

其中,m_k 为 C_k 出现的次数,m_{kj}^{test} 为类别为 C_k 的所有规则样本中第 j 维特征 $X_j^{(test)}$ 出现的次数。但是,可能某些类别在规则样本中没有出现,这样可能会导致 $P(X_j=X_j^{(test)}\mid Y=C_k)=0$,就会影响后验估计。为了解决这种特殊情况,引入了拉普拉斯平滑,可以得到:

$$P(X_j=X_j^{(test)}\mid Y=C_k)=(m_{kj}^{test}+\lambda)/(m_k+O_j\lambda)\tag{11}$$

其中,λ 为一个大于 0 的常数,常取为 1。O_j 为第 j 个特征的取值个数。EasiNB 算法流程如算法 1 所示。

算法 1 EasiNB 规则触发执行预测算法

输入：$\boldsymbol{D}$, k, $\boldsymbol{D}_{test}$

输出：C

1. for $i = 1: m$
2. 初始化规则样本数组 $\boldsymbol{D}_i = (x_1^{(i)}, x_2^{(i)}, \cdots, x_n^{(i)}, y_j)$
3. end for
4. 设定规则执行类别有 k 个类别，分别为 c_1, c_2, …, c_k
5. for $j = 1: k$
6. for $i = 1: m$
7. 分类规则样本，得到 $\boldsymbol{D}_i = (x_1^{(i)}, x_2^{(i)}, \cdots, x_n^{(i)}, y_j)$, y_j $\{c_1, c_2, \cdots, c_k\}$
8. end for
9. end for
10. 提取规则样本数组 $\boldsymbol{D}$ 中的规则特征向量 $\boldsymbol{X}$ 和输出类别 Y
11. 得到规则联合概率分布 $P(X, Y)$
12. 计算规则之间的 PCC
13. for $i = 1: m$
14. if PCC$i \geqslant d$
15. 剔除该样本
16. else 保留在样本空间
17. end for
18. 根据规则的独立性，分解 $P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n \mid Y = C_k)$
19. 计算 $P(Y = C_k)$
20. 计算 $P(X_j = X_j^{(test)} \mid Y = C_k)$
21. 当 $P(Y = C_k \mid X = X^{(test)})$ 取得最大值时，返回规则执行设备控制类别 C
22. 输入新的测试规则样本，根据计算出的控制类别与实际的规则执行设备控制类别是否一致，不一致即视为不安全控制及执行行为

3 实验与性能分析

本节对本文提出的基于监督学习理论的规则触发预测方法 EasiPRED 进行实验验证。选择的应用场景为与实验室合作的煤矿企业，结合《煤矿安全规程(2016)》和煤矿安全监控系统[20]中的气体监测参数以及安全报警措施要求，设置安全控制规则，并根据 1.2 节的规则统一描述标准，转化成标准的规则样本数据集。

规则样本分别以 CH_4、CO、O_2 气体传感器数据和风机、风窗、风门设备的控制信息作为规则的 if 触发条件和 then 执行信息，例如，规则 R 为“O_2 低于 15% 打开风机通风”，O_2 含量低于 15% 作为规则的 if 触发条件，打开风机通风作为规则的 then 执行信息，根据规则统一描述方法，$R = \{10, 1, 5, 0, 1, 15, 100, 1, 10, 1\}$。如表 2 所示，从规则样本数据集中选取了 12 组不同数量的规则样本数据，每组样本的 80% 和 20% 分别作为训练集和测试集。

为了验证 EasiPCA 的降维效果，本文从数据集中随机选择了一组样本数据做结果分析，如图 2 所示。当特征降至一维可保持原特征信息的 54.2%，降至二维可达 93.6%，特征维数降至三维可稳定 99.8% 以上，可以将降维后的 3 个特征值组成的规则特征向量输入到预测模型中做下一步的规则执行控制类别预测。由图3可知，原规则样本中的8个特征值通过降维后，得到3 个新特征值 *char*_1、*char*_2 和 *char*_3，并组成新的规则特征向量。

表 2 规则样本数据集

序号	1	2	3	4	5	6	7	8	9	10	11	12
样本集	100	125	150	200	250	300	350	400	500	600	700	800
训练集	80	100	120	160	200	240	280	320	400	480	560	640
测试集	20	25	30	40	50	60	70	80	100	120	140	160

通过 EasiPCA 得到样本训练集的规则特征向量后，输入至 EasiNB 做模型训练并输出预测的规则执行控制类别。本文的实验评估指标如表 3 所示，其中，TPR 为正确的测试规则与预测为正确的比率，FPR 为错误的测试规则与预测为正确的比率，这两个指标直接关系到规则触发执行后的控制安全，结合准确率 ACC，做了在不同规则样本下的 EasiNB 的预测结果汇总分析，如图 4 所示。

由图 4 可知，当用于训练模型的规则样本数量达到 700 个以上，EasiNB 预测的真正率 TPR 和准确率 ACC 分别稳定于 98%、96% 以上，并且假报率 FPR 可达到0.88% 以下。通过进一步分析可知，对

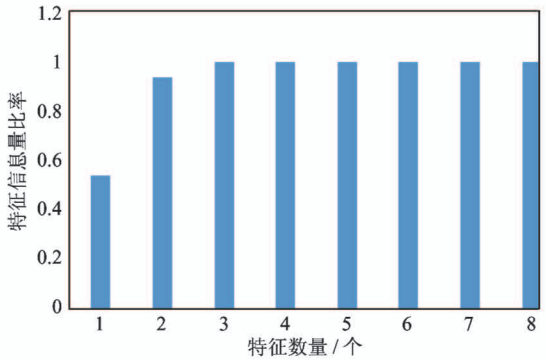

图 2　EasiPCA 特征降维

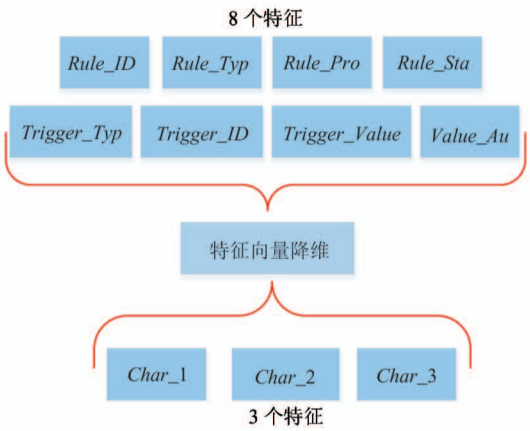

图 3　特征向量降维示意图

表 3　实验评估指标

指标	说明
TP	true positive,对的样本预测为正样本
FP	false positive,错的样本预测为正样本
TN	true negative,对的样本预测为负样本
FN	false negative,错的样本预测为负样本
TPR	ture positive rate,真正率,对的样本被预测为正样本的比率,又称为召回率,TPR = TP/(TP + FN)
FPR	false positive rate,假报率,错的样本被预测为正样本的比率,FPR = FN/(TP + FN)
ACC	预测的准确率

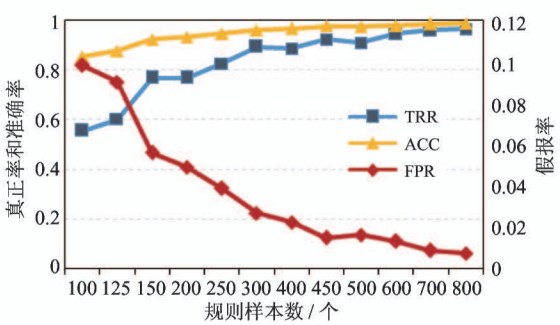

图 4　TPR、FPR、ACC 实验评估汇总

用于训练模型的规则样本集的数量没有太高的要求,少量的样本仍可以得到较高的准确率和真正率,本文选取了第 5 组 300 个规则样本的数据集做预测结果分析,并绘制了标准化混淆矩阵(standardization confusion matrix, SCM),如图 5 所示。由 SCM 矩阵可得到,在样本数量为 300 个的情况下,EasiNB 的预测结果准确率达到了 96%,真正率 TPR 为 89%,假报率仅为 2%。准确率 ACC 与 700 个样本的预测结果一致,而 TPR 有所降低,FPR 由 0.88% 增加到 2%,可以满足对安全控制级别要求不高的场景,如果应用于安全控制级别较高的工业安全监控领域,仍需要增加样本的数量,以提高系统的安全性能。

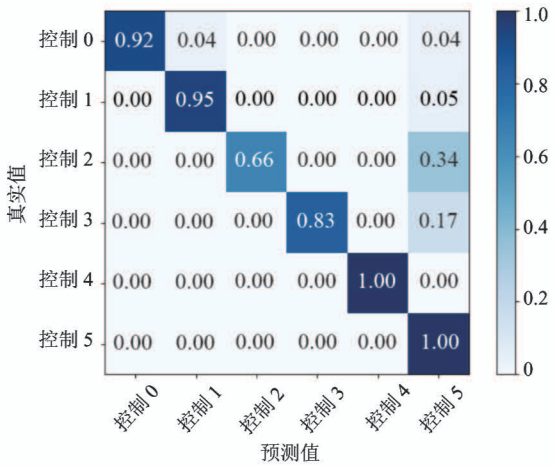

图 5　300 个规则样本的标准化混淆矩阵

为对比 EasiPRED 与其他方法的性能,本文仍选用 300 个规则样本数据集与基于支持向量机(support vector machine, SVM)和决策树(decision tree, DT)的方法进行 ACC、TPR、FPR 对比,如表 4 所示。结果显示,EasiPRED 的 3 项指标均高于其他方法。

表 4　EasiPRED 与其他方法性能对比

方法	EasiPRED	基于 SVM 方法	基于 DT 方法
准确率	0.96	0.94	0.89
真正率	0.89	0.77	0.75
假报率	0.02	0.03	0.03

通过实验可知,给定的规则样本通过 EasiPCA 降维和 EasiNB 预测规则触发执行后的控制类别,并

与实际规则触发执行后的控制类别进行一致性验证,可以识别出实际规则触发执行后的控制行为是否会导致系统发生控制安全事故,能够有效地提高物联网规则引擎在实际应用中的控制安全性能。

4 结 论

针对物联网规则引擎在实际应用中所面临的控制安全问题,本文提出了基于监督学习理论的规则触发执行预测方法 EasiPRED。本文以规则自身为出发点,同时考虑到可能会对系统造成控制安全事故的外部用户、内部用户和系统本身故障的不安全控制因素,通过验证预测规则触发执行后的控制类别与实际规则的控制类别的一致性,避免不安全的控制规则给系统造成无法挽回的安全事故。实验结果表明,EasiPRED 能够有效地识别出不安全的控制规则,提高系统的整体控制安全性能。

未来的研究工作考虑完善规则的统一描述标准,EasiPRED 的实际应用效果首先是取决于规则的正确描述,并且各个应用领域的规则制定标准也不一致,需要为物联网规则引擎的规则制定出台相关标准建立基础,使得物联网规则引擎能够得到进一步的推广和完善。

参考文献

[1] Chernyshev M, Baig Z, Bello O, et al. Internet of Things (IoT): research, simulators, and testbeds[J]. *IEEE Internet of Things Journal*, 2018, 5(3): 1637-1647

[2] 池深启. 轻量级规则引擎 Drools 在物联网平台中的应用研究[D]. 杭州:浙江工业大学电子与通信工程学院, 2017: 5-15

[3] Zhang J D, Yang J X, Li J. When rule engine meets big data: design and implementation of a distributed rule engine using spark[C]//IEEE 3rd International Conference on Big Data Computing Service and Applications, San Francisco, USA, 2017: 41-49

[4] Roy K S, Kalita H K. A survey on authentication schemes in IOT[C]//IEEE International Conference on Information Technology, Bhubaneswar, India, 2017: 335-339

[5] Wang Q, Gao M Z, Qu G. PUF-PassSE: a PUF based password strength enhancer for IoT applications[C]// IEEE International Symposium on Quality Electronic Design, Santa Clara, USA, 2019: 510-516

[6] 彭昆仑. 物联网中轻量安全身份可信机制研究与实现[D]. 合肥:国防科学技术大学研究生院, 2016: 27-50

[7] Shah T, Venkatesan S. Authentication of IoT device and IoT server using secure vaults[C]//IEEE International Conference On Trust, Security And Privacy In Computing And Communications, New York, USA, 2018: 819-824

[8] Das A K, Wazid M, Yannam A R, et al. Provably secure ECC-based device access control and key agreement protocol for IoT environment[J]. *IEEE Access*, 2019, 7: 55382-55397

[9] Chen H, Chang C, Leu F. Implement of agent with role-based hierarchy access control for secure grouping IoTs[C]//IEEE Annual Consumer Communications and Networking Conference, Las Vegas, USA, 2017:120-125

[10] Wang H, Wang H, Yue H, et al. Prediction of user behavior in smart home based on improved ARIMA model[C]//IEEE International Conference on Mechatronics and Automation, Changchun, China, 2018:298-302

[11] Wu E, Zhang P, Lu T, et al. Behavior prediction using an improved hidden Markov model to support people with disabilities in smart homes[C]//IEEE 20th International Conference on Computer Supported Cooperative Work in Design, Nanchang, China, 2016:560-565

[12] Dang T, Tran M, Le D, et al. Trend-adaptive multi-scale PCA for data fault detection in IoT networks[C]//International Conference on Information Networking, Chiang Mai, Thailand, 2018:744-749

[13] Kumar D, Smys S, Smilarubavathy G, et al. Fault detection methodology in wireless sensor network[C]//The 2nd International Conference on I-SMAC, Palladam, India, 2018:723-728

[14] 田瑞琴, 吴尽昭, 唐鼎. 物联网网关中轻量化规则引擎的设计与实现[J]. 计算机应用, 2015, 35(4): 1035-1039

[15] 汤冬冬. 基于 Drools 的分布式业务规则引擎的设计与实现[D]. 大连: 大连理工大学软件学院, 2010: 4-12

[16] 黄晓辉, 李栋, 石海龙, 等. EasiRCC:一种针对智能家居的规则匹配和冲突消除方法[J]. 计算机研究与发展, 2017, 54 (12): 2711-2720

[17] 赵博轩, 房宁, 赵群飞, 等. 利用拼音特征的深度学习文本分类模型[J]. 高技术通讯, 2017, 27(7):

596-603

[18] 王金甲, 陈浩, 刘青玉. 大数据下的深度学习研究[J]. 高技术通讯, 2017, 27(1): 27-37

[19] Zhang X Y, Shi H C, Zhu X B, et al. Active semi-supervised learning based on self-expressive correlation with generative adversarial networks [J]. *Neurocomputing*, 2019, 345: 103-113

[20] 中华人民共和国应急管理部. AQ6201-2019 煤矿安全监控系统通用技术要求[S]. 北京:应急管理出版社, 2019

Research on the rule trigger execution prediction method based on supervised learning

Huang Xiaohui[*,**], Cui Li[**], Huang Xi[**]

([*] University of Chinese Academy of Sciences, Beijing 100049)

([**] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

Abstract

Aiming at the control and execution security problems caused by external and internal users and the system itself in the practical application of Internet of Things (IoT) rule engine, a control prediction method of rule trigger execution device (EasiPRED) based on supervised learning theory is proposed. First of all, this paper has developed the unified description standard of Internet of Things rule engine, which is convenient for the establishment and management of rules. Then, a feature extraction algorithm based on principal component analysis (PCA) is proposed to reduce the dimensionality of rule sample sets and extract rule features, which not only preserves the original feature information, but also reduces the computational complexity. Finally, based on the supervised learning method, the rule sample set is trained and learned, the control category of the rule is predicted after the execution is triggered, and the consistency of the predicted result and the output of the actual rule is verified, so as to judge the safety of the control behavior. Experimental results show that EasiPRED can effectively identify unsafe control rules, improve the overall safety performance of the system, and avoid control and execution safety accidents.

Key words: rules engine of Internet of Things (IoT), insecurity factor, control security, monitor learning, control category