

# 基于贝叶斯网络 EM 算法模型的工控蜜罐识别

张立芳<sup>1</sup>, 王 钢<sup>2</sup>, 颜培志<sup>2</sup>, 姚 旭<sup>1</sup>, 孙 叶<sup>1</sup>

(1. 内蒙古工业大学 信息工程学院, 内蒙古 呼和浩特 010051;

2. 内蒙古工业大学 信息化建设与管理中心, 内蒙古 呼和浩特 010051)

**摘 要:**随着工控设备越来越多暴露于互联网,面临的安全威胁不断增加,主动防御已经成为一种必要的防御手段,蜜罐技术是一种有效的主动防御技术。攻击者为了攻击真实的资产设备,研究人员开始研究识别蜜罐的方法。对蜜罐进行准确识别涉及到许多不确定性因素。贝叶斯网络用于解决不确定性问题,与蜜罐识别问题相符合。基于蜜罐识别与贝叶斯网络的特点,提出了贝叶斯网络参数学习 EM 算法模型的工控蜜罐识别方法。首先,介绍了贝叶斯网络的理论基础及贝叶斯网络用于蜜罐识别的优势;接着,描述参数建模所用算法及预测推理算法,完成用于识别蜜罐的贝叶斯网络模型;最后,通过与 SVM、KNN、随机森林和 Native bayes 算法作对比实验,验证所采用贝叶斯网络 EM 算法训练模型的性能更优,该模型借助贝叶斯联结树推理算法来完成预测识别,通过实例分析进行验证。实验结果表明,用 EM 算法训练的模型对于识别蜜罐是有效的。

**关键词:**贝叶斯网络;蜜罐识别;参数建模;推理;预测

**中图分类号:**TP301.6

**文献标识码:**A

**文章编号:**1673-629X(2022)08-0116-06

doi:10.3969/j.issn.1673-629X.2022.08.019

## Industrial Control Honeypot Recognition Based on Bayesian Network EM Algorithm Model

ZHANG Li-fang<sup>1</sup>, WANG Gang<sup>2</sup>, YAN Pei-zhi<sup>2</sup>, YAO Xu<sup>1</sup>, SUN Ye<sup>1</sup>

(1. School of Information Engineering, Inner Mongolia University of Technology, Hohhot 010051, China;

2. Information Construction and Management Center, Inner Mongolia University of Technology, Hohhot 010051, China)

**Abstract:**As more and more industrial control equipment are exposed to the Internet, the security threats are increasing. Active defense has become a necessary defense method, and honeypot technology is an effective active defense technology. In order to attack the real asset equipment, the researchers have begun to study the method of identifying the honeypot. Accurate identification of honeypot involves many uncertainties factors. Bayesian network is used to solve the uncertainty problem, which is consistent with the honeypot identification problem. Based on the characteristics of honeypot recognition and Bayesian network, we propose an industrial control honeypot recognition method based on Bayesian network parameter learning EM algorithm model. Firstly, the theoretical basis of Bayesian networks and the advantages of Bayesian networks for honeypot identification are introduced. Then, the algorithms used in parameter modeling and predictive inference algorithms are described, and the Bayesian network model for honeypot identification is completed. Finally, by comparing the experiments with SVM, KNN, Random Forest and Native Bayes algorithm, the performance of the training model of Bayesian network EM algorithm is verified, and the prediction recognition is accomplished by the Bayesian junction tree reasoning algorithm, which is verified through case analysis. Experimental results show that the model trained with the EM algorithm is effective for identifying honeypot.

**Key words:**Bayesian network;honeypot recognition;parameter modeling;reasoning;prediction

## 0 引 言

工业互联网时代下,大数据、人工智能的发展使得越来越多原本处于孤立环境中的工业控制设备暴露于公共互联网,遭到来自互联网的攻击威胁<sup>[1]</sup>。工业控

制系统(ICS)的安全问题变得越来越普遍,蜜罐和反蜜罐作为 ICS 安全的重要组成部分已经成为攻防对抗的重点,随着蜜罐技术的不断完善发展,针对蜜罐的识别技术也在相应的提高。工业控制系统蜜罐技术<sup>[2]</sup>作

收稿日期:2021-09-16

修回日期:2022-01-18

基金项目:内蒙古自治区教育基金(NJZZ18077)

作者简介:张立芳(1996-),女,硕士研究生,研究方向为工控安全、贝叶斯网络;王 钢,正高级工程师,研究方向为网络信息安全。

为主动诱捕手段之一,能够有效捕获针对工业控制系统发起的网络攻击,保护真实工控资产设备。安全研究者为了完善蜜罐技术,开始从攻击者角度出发,研究蜜罐识别技术<sup>[3]</sup>。

传统蜜罐识别都是针对单一特征,Sebek 是一种基于内核的数据捕获机制,常用于构建高交互蜜罐,因此,识别出 Sebek 机制就可以确定目标设备是蜜罐。朱一帅<sup>[4]</sup>针对 Sebek 机制进行研究与分析,对于识别蜜罐有重要意义。近几年,人工智能和机器学习的广泛应用,使得研究人员开始从这个角度研究蜜罐识别技术,北京邮电大学的程卓提出基于随机森林模型的工业控制系统蜜罐识别,解决了单一特征识别的局限性<sup>[5]</sup>。从目前的情况看,对工控蜜罐的识别技术仍相对教少,大部分研究的是对蜜罐的识别,因此该文提出专门针对工控蜜罐的识别方法。

从目前了解的研究情况看,没有专家将贝叶斯网络用于工业控制蜜罐识别的研究上,针对蜜罐捕获的攻击数据及数据的分析问题,有研究人员采用动态贝叶斯网络的方法,分析蜜罐获取的数据,根据贝叶斯的递推公式对攻击行为进行预测,并将它应用于数据的分析<sup>[6]</sup>。攻击类型有很多种,在面对未知攻击时,有学者提出采用贝叶斯网络推理算法分析和判断蜜罐中的未知攻击<sup>[7]</sup>。贝叶斯与蜜罐结合使用能够解决一些实际问题。工控蜜罐识别面临不确定性,与贝叶斯网络应用问题相符合。作为不确定性分析的重要工具,用贝叶斯网络进行蜜罐识别可行,因此将贝叶斯网络运用于工控蜜罐识别研究上很有意义。

## 1 贝叶斯网络介绍

### 1.1 基本概念

贝叶斯网络是一种基于概率的不确定性推理方法,对不确定性问题具有强大的处理能力和自我学习更新能力。贝叶斯网络是 Judea Pearl 于 1988 年提出的一种基于概率推理的图形化网络<sup>[8]</sup>,它将若干具有因果关系或概率相关性的事件以网络形式表示出来,然后在不同的事件中根据先验概率进行推理计算,获得各种事件发生的概率值。

贝叶斯网络是一个有向无环图(DAG),由代表变量节点及连接这些节点的有向边构成。节点代表随机变量,节点间的有向边代表了节点间的因果关系,用条件概率表达关系强度<sup>[9]</sup>,条件概率表(conditional probability table, CPT)是反映变量之间关联性的局部概率分布,即概率参数,适用于表达和分析不确定性和概率性的事件。

在一个随机试验中,有  $n$  个互相排斥、独立的事件

$A_1 \cdots A_n$ , 如果  $P(A_i)$  表示事件  $A_i$  发生的概率,且

$\sum_{i=1}^n P(A_i) = 1$ ,  $B$  记为任一事件,则有:

$$P(A_i | B) = \frac{P(B | A_i)P(A_i)}{\sum_{j=1}^n P(B | A_j)P(A_j)}, i = 1, 2, \cdots, n \quad (1)$$

公式(1)就是著名的贝叶斯公式。其中  $P(A_1), P(A_2), \cdots, P(A_n)$  称为先验概率。事件  $B$  发生时,由于这个新情况的出现,对于事件  $A_1 \cdots A_n$  发生的可能性有了新的认识,即在事件  $B$  发生的条件下事件  $A$  发生的情况称为事件  $A$  的后验概率。综合了先验信息和提供的新信息,形成了关于  $A_i$  发生概率的新认识。这个由先验信息到后验信息的转化,是贝叶斯统计的特征。

### 1.2 贝叶斯网络用于蜜罐识别的优势

贝叶斯方法基于概率推理,用来解决不确定性问题,与该文识别蜜罐所面临的不确定性一致。因此,贝叶斯网络用于蜜罐识别具有一系列的优势<sup>[10]</sup>:①贝叶斯网络的评估结果不仅反映了当前的信息,而且综合了历史和先验知识,能够更加准确地预测蜜罐;②贝叶斯网络能够处理各种不确定性信息和不完备数据集,与完成蜜罐识别所采用数据集相符;③贝叶斯网络与一般表示方法不同的是对于问题域的建模,当条件或行为等发生变化时,不用对模型进行修正;④采用贝叶斯网络参数建模 EM 算法和贝叶斯联结树推理算法,提高了蜜罐识别准确率。

## 2 参数建模与推理算法

### 2.1 数据

智能的蜜罐识别技术方法的主要挑战是数据,Shodan<sup>[11]</sup>是著名的网络空间搜索引擎,用于搜索网络设备,这个平台已经标记了许多已知的蜜罐服务器以及大量的工业控制系统资产设备,从平台上获取蜜罐及工控系统数据作为数据集,用来构建 BN(贝叶斯网络)。该数据库包含各种用来判断蜜罐的特征属性(port、serial number of module、PLC name、module name),这些数据作为实验研究的数据集是可靠的。除了从 Shodan 下载的数据之外,还部署了一些工控蜜罐,利用 Nmap 扫描工具对部署的蜜罐进行脚本扫描,获取一些 OS 的指纹数据。

从 Shodan 下载的数据为 json 格式,为方便完成 BN 建模,一般需要将数据格式转换为 csv。依据工控蜜罐特征分析提取相关特征列,使与贝叶斯网络模型节点相对应,作为模型训练和测试数据集。该文收集了 1 053 条数据,蜜罐记录数量为 307,实际设备数量为 746,数据集如表 1 所示。

表 1 实验数据集

Data type	Size
Honeypot records	307
Real system records	746

## 2.2 参数建模

### 2.2.1 贝叶斯网络结构搭建

基于贝叶斯网络进行蜜罐识别,需按照一定的方法和原则,构建一个合理的网络,主要有以下 3 个步骤<sup>[12]</sup>:①确定节点,贝叶斯网络由节点组成,节点对应不同的事件,首先必须确定存在哪些可以识别为蜜罐的特征,即确定出现哪些特征可以判定是蜜罐。②确定节点关系即事件之间的因果关联,由于选定的节点均为蜜罐特征,它们的出现将用来判断是否为蜜罐,因果关系明确。③概率分配,对于没有父节点的事件指定先验概率,即  $P(A_i)$ ;对于有父节点的事件指定条件概率,即  $P(A_i | \text{pa}(A_i))$ 。

根据以上 3 个步骤,很容易构建出贝叶斯网络结构模型。

### 2.2.2 参数学习 EM 算法

参数学习的定位角色是在已知初始化模型的基础上,包括初始化的结构和初始化的参数,基于实时获取的新数据,通过参数学习来优化更新模型,解决实时性问题。通常贝叶斯网络中变量的许多概率分布是未知的,希望从数据(即通过实验、文献或其他来源获得的一系列观察数据)中了解这些概率(参数)。一种被称为 EM(估计最大化)算法的算法对这种参数学习特别有用。EM 试图从观测(但往往是不完整的)中找到网络的模型参数(概率分布)。适用于给定贝叶斯网络的结构和样本数据,在已知先验的情况下,根据贝叶斯推理计算模拟样本缺失的概率,利用计算所得的期望补全缺失的数据集,重新对当前的网络参数进行学习。一般收集的实验数据存在缺值,EM 算法用来解决数据不完整的参数估计问题,选择 EM 算法作为缺值数据的 BN 参数学习算法是合适的。

EM 算法由两个步骤组成,求期望的 E 步和求最大似然估计的 M 步<sup>[13]</sup>。E 步骤:根据参数初始值或上一次迭代的模型参数计算隐变量的后验概率即隐变量的期望。

$$Q_i(z^{(i)}) = P(z^{(i)} | x^{(i)}, \theta) \quad (2)$$

M 步骤:将似然函数最大化以修正新的参数值:

$$\theta = \arg \max_{\theta} \sum_i \sum_{z^{(i)}} Q_{(i)}(z^{(i)}) \log \left( \frac{P(x^{(i)}, z^{(i)}, \theta)}{Q_{(i)}(z^{(i)})} \right) \quad (3)$$

不断的迭代,就可以得到使似然函数  $L(\theta)$  最大化的参数  $\theta$ 。

## 2.3 推理

### 2.3.1 推理算法

联结树算法是 Hugin Expert 工具默认的推理算法,是一种精确的推理算法,目前速度最快。贝叶斯网络(BN)是在联结树的二级结构(SS)中进行推理的<sup>[14]</sup>。

联结树推理算法的基本思想是,将 BN 转化为一种 SS,再通过对 SS 推理得到 BN 推理的精确结果。二级结构  $SS = (JT, PP)$ ,  $JT = (C, S)$  为联结树,  $C$  为 BN 中的团(clique)也是联结树的节点,联结树的节点之间的连接为  $S$ ,称为 JT 的边,是收集证据和分配证据的通信通道。PP 为团和边相关的概率势(probability potential),从每个团中变量的联合概率分布计算得到。

联结树推理算法基本步骤为:

①将贝叶斯网络转化为联结树;找出 BN 中每一个节点的父节点,并将它们用无向边两两相连,同时将所有 BN 的有向边改为无向边,建立 BN 的 Moral 图,在 Moral 图中添加一些无向边,将每一个等于或大于 4 的环的两个非相邻节点连接起来,完成 Moral 图的三角化。对三角化后的 Moral 图,找到构成联结树的所有团。在找到的团中添加一些边构造一棵联结树,树中连接任意两个团的边的所有团节点必须包含两个团节点的交集。

②初始化;为联结树的所有节点指定参数。

③消息传递;通过各团节点之间的消息传递,使联结树达到稳态。向内消息传递,消息从联结树的叶子发送到树的根部,即收集证据。向外消息传递,其中消息从树的根部向叶子发送,即分配证据。

④概率计算;找到任意一个包含变量  $V$  的团节点  $C$ ,通过公式(4)计算变量  $V$  的分布。

$$P(V) = \sum_{C \setminus \{V\}} \tau_C \quad (4)$$

其中,  $\tau_i$  代表  $C_i$  的分布函数。

⑤加入证据;在新的证据加入时,要重新收集证据和分配证据,直到联结树达到稳态。对任意的团节点  $C$  有:

$$\tau_C = P(C, e) \quad (5)$$

其中,  $e$  表示加入的证据。计算变量  $V$  的概率分布,首先找到任意一个包含变量  $V$  的团节点  $C$ ,计算公式(6),再根据条件概率公式,计算变量  $V$  的概率分布。

$$P(V | e) = \sum_{C \setminus \{V\}} P(C, e) = \sum_{C \setminus \{V\}} \tau_C \quad (6)$$

### 2.3.2 推理过程

BN 结构固定,父节点是一系列特征,子节点是蜜罐特征。BN 节点特征和数据集的属性对应,在对数分析后,通过出现的具体特征状态来预测蜜罐概率。

BN 模型的目标是通过特征来预测判断蜜罐的概率大小,作为攻击方,通过攻击手段获取到目标设备的信息,根据获取的特征,使用建模完的模型进行预测。当 BN 切换到运行模式,激活自带的推理算法—联结树算法,通过输入证据,调用算法得出预测概率。

BN 的结果标签 honeypot 有 1 和 0 两种状态,在没有输入证据之前,honeypot 标签的状态概率固定。当输入特征证据,证据通过消息传递,调用联结树算法,计算相应 honeypot 标签状态概率。当有多个特征证据同时输入,经过联结树算法的推理,计算出最终 honeypot 标签状态概率,honeypot 为 1 的状态是预测为蜜罐的概率。

### 3 贝叶斯网络模型用于蜜罐识别实例分析

#### 3.1 蜜罐识别 BN 的构建与分析

蜜罐基于爬虫技术,具有良好的欺骗性能,蜜罐识别需要考虑区别于 ICS 设备的多个特征。该文考虑的主要特征包括:端口号、设备的串行序列号、系统名、设备名称。根据主要蜜罐特征结合以上网络构建的步骤,完成用于蜜罐识别的贝叶斯网络模型,如图 1 所示。

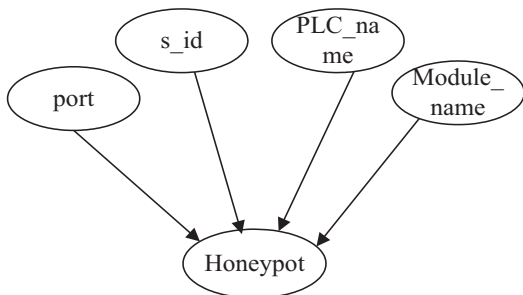


图 1 蜜罐识别贝叶斯网络模型

其中:port 为端口,PLC\_name 为系统名,s\_id 为设备串行序列号,Module\_name 为设备名称,Honeypot 为蜜罐。模型中的变量状态集合如下:

端口 (port): port = 102、502、21、47808、1962、20547、2002

系统名 (PLC\_name): PLC\_name = Technodrome、ET 200S station\_1、12172306、GDV570、SIMATIC 300 (1)、SIMATIC 300、SAAP7-SERVER、BASE DIR R61. 2、VENT、CPU 314C、Production Management、H、Central Pump、S7300/ET200M station\_2

设备串行序列号 (s\_id): s\_id = 88111222、100194、S C - C2UR28922012、S Q - DNU118252013、S C - BOVM84702011、S C - E4UL21922014、S C - A6TD27832010、S C - W8V004032008、S C - D8U562202013

设备名称 (Module\_name): Module\_name =

Siemens, SIMATIC, S7-200、CPU 314C-2 DP、CPU 314C-2 PN/DP、CPU 315-2 PN/DP、CPU 315-2 DP、CPU 314、TriKantel、Energy, Water, Climate C、Pump Control Unit、PLC\_1。

蜜罐 (Honeypot): Honeypot = 1、0

在实际应用中,每个特征会有多种取值,该文从简单的角度出发,只选取了所用数据集中的一些状态特征。

#### 3.2 节点概率分配

网络构造完成,下一个任务就是生成条件概率表。这些概率可以通过专家的经验获得,也可以通过计算机对原有数据进行统计学习获得,或是两者的有机结合。本例中需要指定的先验概率包括  $P(\text{port})$ 、 $P(s\_id)$ 、 $P(\text{PLC\_name})$ 、 $P(\text{Module\_name})$ ,由于没有专家经验,设 4 类特征的状态出现的可能性均等,例如表 2 中 s\_id 的先验概率。

表 2 s\_id 特征的先验概率

s_id 特征状态	先验概率	后验概率
88111222	0.111 11	0.718 74
100194	0.111 11	0.015 62
S C - C2UR28922012	0.111 11	0.140 62
S Q - DNU118252013	0.111 11	0.015 62
S C - BOVM84702011	0.111 11	0.015 62
S C - E4UL21922014	0.111 11	0.046 87
S C - A6TD27832010	0.111 11	0.015 62
S C - W8V004032008	0.111 11	0.015 62
S C - D8U562202013	0.111 11	0.015 62

已知各节点先验概率和条件概率后,使用搭建好的结构模型采用 EM 参数学习算法对各节点参数更新,经过多次迭代,参数最终保持不变。参数建模完成后,4 类特征的状态参数改变,对应 s\_id 特征状态的参数如表 2 中后验概率所示。

#### 3.3 评价方法

模型评估是实验成败的关键,目的是评估最终模型的准确程度。文中采用精确率、召回率、F1 score、ROC 曲线和 AUC 值作为评价指标,如公式 (7) 到公式 (9) 所示。

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

$$\text{F1 score} = \frac{2TP}{2TP + FP + FN} \quad (9)$$

其中,TP 为预测为蜜罐真实也为蜜罐的数量;TN 为预测为非蜜罐真实也为非蜜罐的数量;FP 为预测为蜜罐真实为非蜜罐的数量;FN 为预测为非蜜罐真实为蜜罐

的数量。

召回率为正确分类为蜜罐的所有样本与真实为蜜罐的所有样本比例;准确率为正确分类为蜜罐的所有样本与预测为蜜罐的所有样本比例。ROC 由 TPR 和 FPR 构成,TPR 是正确分类为蜜罐的所有样本与真实为蜜罐的所有样本的比例,如公式(10)所示;FPR 是错误预测为蜜罐的样本与真实为非蜜罐所有样本的比例,如公式(11)所示。ROC 曲线中,X 轴为 FPR,Y 轴为 TPR。

$$TPR = \frac{TP}{TP + FN} \quad (10)$$

$$FPR = \frac{FP}{FP + TN} \quad (11)$$

ROC 曲线在(0,0)和(1,1)之间的对角线上方,表明模型的性能是可以接受的,为了更直观评估模型,研究人员通常使用 AUC 指标来衡量模型的整体效率,AUC 为 ROC 曲线下的面积,AUC 的值越接近 1,模型的效果越好。

### 3.4 实验

研究贝叶斯网络 EM 算法训练模型的模型性能与其他几种机器学习算法训练模型性能的对比优势,和贝叶斯推理算法对预测蜜罐结果的准确性。在实验中,使用 scikit-learn 库完成机器学习模型的训练,计算过程在 i5-7200 CPU,12 GB 内存的计算机上运行。EM 算法模型训练借助 Hugin 贝叶斯工具。

#### 3.4.1 对比实验

未经比较的结果是不可靠的,文中设计了这种比较实验,以突出文中方法的优势。回顾一些机器学习的研究<sup>[15]</sup>,选择了 SVM、KNN、随机森林和 Native bayes 作为贝叶斯参数学习 EM 算法的比较对象。将对比的所有机器学习算法用在同一数据集上分别训练 4 个模型,对每个模型列出它们的精确率、召回率、准确率和 F1 score,且在同一坐标系中绘制它们各自的 ROC 曲线,模型的性能一目了然。文中采用的算法也在相同的数据集上进行训练,列出精确率、召回率、准确率、F1 score 且单独画出 ROC 曲线,对比 4 种机器学习算法。

#### 3.4.2 实验结果

表 3 是 4 种机器学习算法和 EM 算法在同一数据集下所训练模型的召回率、精确率、准确率和 F1 score。图 2 是四种机器学习算法的 ROC 曲线。图 3 是文中采用算法的 ROC 曲线。图 2 虚点线是随机森林算法的 ROC 曲线,AUC 值为 0.955 6,实线是 SVM 算法的 ROC 曲线,AUC 值为 0.933 3,点线是 KNN 算法的 ROC 曲线,AUC 值为 0.933 3,虚线是 Native bayes 算法的 ROC 曲线,AUC 值为 0.955 6。图 3 是文

中算法借助工具完成的 ROC 曲线图。通过两图对比,对于解决蜜罐识别所用的数据集,明显文中采用的 EM 算法效果更好,AUC 值为 0.963 8。表 3 中,文中所采用算法训练模型的召回率是 0.979,准确率是 0.979,精确率是 0.97,F1 score 是 0.979,对比其他算法的模型指标性能更好。使用贝叶斯参数学习 EM 算法所训练模型的 ROC 曲线和模型评估指标证明了最终模型具有较高的检测率和良好的泛化能力。

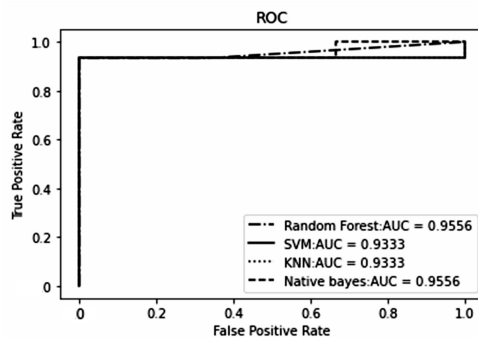


图 2 四种机器学习算法的 ROC 曲线

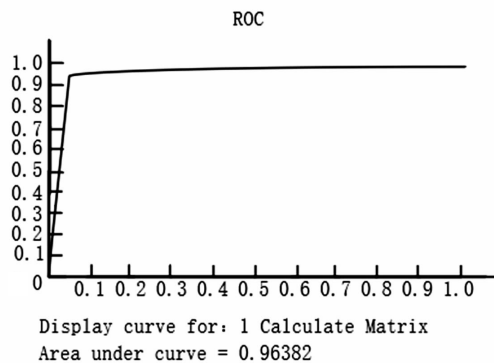


图 3 贝叶斯网络参数学习 EM 算法 ROC 曲线  
表 3 模型的召回率、精确率、F1 score、准确率

算法	Recall	Precision	Accuracy	F1 score
随机森林	1	0.933	0.952	0.966
SVM	0.933	0.933	0.905	0.933
KNN	1	0.933	0.952	0.966
Native bayes	1	0.933	0.952	0.966
EM	0.979	0.979	0.970	0.979

根据 AUC 值的判断标准,AUC 值越接近 1,模型的效果越好。文中采用 EM 算法模型的 AUC 值为 0.963 8,高于其他机器学习算法训练的模型,模型性能较好。

### 3.5 预测

完成参数建模及模型评估后,使用联结树推理算法完成贝叶斯网络的计算推理。参数建模后的贝叶斯网络模型固定,借助贝叶斯推理算法进行蜜罐的识别预测。假设输入证据 s\_id 状态为 88111222,经过联结树推理算法的计算,此时 honeypot 标签状态为 1 的概

率由原来的 62.7% 变为 67.91%,说明当出现此特征状态时,预测为工控蜜罐的概率是 67.91%。同时输入证据 s\_id 状态为 88111222 和 PLC\_name 状态为 Technodrome,此时 honeypot 标签状态为 1 的概率由原来的 62.7% 变为 79.1%。当同时出现这两个状态时,预测为工控蜜罐概率的可能性增加。

根据获取到的特征状态,来预测工控蜜罐的概率。判断是否是蜜罐有一个阈值,在完成参数建模之后,此时的 honeypot 标签状态为 1 的概率是 62.7%,以此值作为阈值。当输入证据后,通过推理算法计算得出 honeypot 状态为 1 的输出概率值大于 62.7%,预测蜜罐的概率为输出值;当输入证据后,计算得出的输出概率值小于 honeypot 状态为 1 的阈值时,判断不是蜜罐。概率值的大小反应判定是蜜罐的可能性。表 4 所示为一些蜜罐的预测概率结果。

表 4 蜜罐识别的预测概率

evidence	P(honeypot)=1
s_id(88111222)	0.679 1
PLC_name( Technodrome)	0.709 2
s_id(88111222) 和 PLC_name( Technodrome)	0.791 0
s_id(100194)	0.499 2
Module_name( Siemens, SIMATIC, S7-200)	0.702 5
port(20256)	0.485 1

#### 4 结束语

文中提出一种基于贝叶斯网络 EM 算法的工控蜜罐识别方法,该方法首先使用 Shodan 上收集的数据,采用 EM 算法训练稳定的模型,然后基于 Hugin 自带的推理算法完成预测识别。贝叶斯网络对于不确定性事件的概率推测,与该文对蜜罐识别的不确定性问题相符合,对于处理识别蜜罐的不确定性具有特殊的优势。作为攻击方对于目标系统或设备是否是蜜罐未知,结合一些蜜罐特征使用贝叶斯网络模型预测出现某个特征导致是蜜罐的概率,以便更加精确地识别蜜罐。相对于预测为蜜罐概率是 100%,67.91% 更准确,因为本质上并不知道目标设备是否确定为蜜罐。

实验结果表明,对比其他模型,文中采用的 EM 算法模型性能更优秀。基于贝叶斯网络模型结合贝叶斯推理算法来预测,提高了蜜罐识别的准确性。

#### 参考文献:

- [1] 范文斌. 工业控制协议安全防护分析[J]. 电子科学技术, 2015, 2(3): 334-337.
- [2] IRVENE C, FORMBY D, LITCHFIELD S, et al. HoneyBot: a honeypot for robotic systems[J]. Proceedings of the IEEE, 2018, 106(1): 61-70.
- [3] KRAWETZ N. Anti-honeypot technology[J]. IEEE Security & Privacy, 2004, 2(1): 76-79.
- [4] 朱一帅, 吴礼发. 基于 Sebek 的蜜罐识别机制研究[J]. 信息技术, 2009, 33(1): 83-86.
- [5] 陈卓. 基于无状态连接的工控系统扫描平台的设计与实现[D]. 北京: 北京邮电大学, 2018.
- [6] 鲍巍, 于博. 动态贝叶斯网络在蜜罐上的研究与应用[J]. 计算机安全, 2012(12): 46-49.
- [7] 鲍巍, 黄振颖. 贝叶斯网络在蜜罐系统中的应用研究[J]. 网络安全技术与应用, 2013(4): 36-38.
- [8] PEARL J. Fusion, propagation, and structuring in belief networks[J]. Artificial Intelligence, 1986, 29(3): 241-288.
- [9] KJRULFF U B, MADSEN A L. Bayesian networks and influence diagrams: a guide to construction and analysis[M]. New York: Springer Science+Business Media, 2013.
- [10] 王朔, 周少平, 黄教民. 基于贝叶斯网络的威胁识别[J]. 计算机工程与设计, 2006, 27(18): 3442-3443.
- [11] 连晓伟. 基于蜜罐技术的 Shodan 扫描特征研究[D]. 太原: 太原理工大学, 2020.
- [12] 李曼, 冯新喜. 基于贝叶斯网络的威胁识别[J]. 现代防御技术, 2009, 37(5): 10-13.
- [13] 张少中, 章锦文, 张志勇, 等. 面向大规模数据集的贝叶斯网络参数学习算法[J]. 计算机应用, 2006, 26(7): 1689-1691.
- [14] 胡小建, 杨善林, 马溪骏. 基于联结树的贝叶斯网的推理结构及构造算法[J]. 系统仿真学报, 2004, 16(11): 2559-2563.
- [15] HUANG C, HAN J, ZHANG X, et al. Automatic identification of honeypot server using machine learning techniques[J]. Security and Communication Networks, 2019: 1-8.