基于机器学习的 Modbus_TCP 通信异常检测方法研究*

陈鑫龙[1],陈志翔[1,2],周小方[2,3]

(1.闽南师范大学 计算机学院,福建 漳州 363000;
2.数据科学与智能应用福建省高校重点实验室,福建 漳州 363000;
3.闽南师范大学 物理与信息工程学院,福建 漳州 363000)

摘 要:针对工业控制系统中 Modbus_TCP 协议存在的诸多安全隐患问题,提出了基于机器学习的 Modbus_TCP 通信异常检测方法,分析了 Modbus_TCP 报文类型与结构特点,介绍了机器学习中决策树分类模型算法的实现过程,建立了 Modbus_TCP 协议的模拟通信,使用了 Scapy 工具构造伪报文实现异常检测。设置了朴素贝叶斯分类模型、逻辑回归分类模型和传统支持向量机分类模型的实验与之对比,并且对模型的准确率、误报率、漏报率和时间性能进行分析。分析结果表明,决策树分类模型准确率高,消耗时间短,具有一定的优越性。

关键词:Modbus_TCP 协议;工业控制系统;决策树算法;异常检测

中图分类号:TP309 **文献标识码**:A **DOI**:10.19358/j.issn.2096-5133.2020.10.011

引用格式:陈鑫龙,陈志翔,周小方. 基于机器学习的 Modbus_TCP 通信异常检测方法研究[J].信息技术与网络安全,2020,39(10):55-60.

Research on Modbus_TCP communication anomaly detection method based on machine learning

Chen Xinlong[1], Chen Zhixiang[1,2], Zhou Xiaofang[2,3]

(1.School of Computer Science, Minnan Normal University, Zhangzhou 363000, China;
2.Key Laboratory of Data Science and Intelligence Application, Zhangzhou 363000, China;
3.School of Physics and Information Engineering, Minnan Normal University, Zhangzhou 363000, China)

Abstract: Aiming at the hidden security problems of Modbus_TCP protocol in industrial control systems, this paper proposes a Modbus_TCP communication anomaly detection method based on machine learning, analyzes the types and structural characteristics of Modbus_TCP messages, introduces the implementation process of decision tree classification model algorithm in machine learning, establishes the simulation communication of Modbus_TCP protocol, and uses Scapy tool to construct pseudo message to realize anomaly detection. The experiments of Naive Bayes classification model, logistic regression classification model and traditional support vector machine classification model are also set up to compare with the proposed method, and the accuracy, false positive rate, false negative rate and time performance of the models are analyzed. The analysis results show that the decision tree classification model has high accuracy, short time consumption, and certain advantages.

Key words: Modbus_TCP protocol; industrial control system; decision tree algorithm; anomaly detection

0 引言

随着两化融合进程的不断加速,工业控制系统逐渐接入互联网,使得原本的“工业信息孤岛”变得不再封闭,但同时也不再安全。近几年,全球工控安全事件频发,不仅带来了巨大的经济损失,同时也给人们的生活环境及人身安全带来了巨大的影响。Modbus 协议是工业控制系统(Industrial Control System, ICS)中的一种常用的通信协议,其具有实现

* 基金项目:福建省中青年教师教育科研项目(JAT191414);漳州市自然科学基金(ZZ2020J33);闽南师范大学研究生教改项目(MSYJG 8)

简单、部署方式多样、标准公开等诸多优势，但同时也存在缺乏认证机制、授权机制、加密机制和功能码滥用等诸多缺陷，给系统带来了一定的安全威胁。

国内外许多专家学者对这一领域进行了研究，EREZ N 等人提出了基于有限自动机(Deterministic Finite Automaton，DFA)算法的异常检测模型[1]，该方法将数据包的每个字段都作为样本特征，进行深度检测，虽然可以有效地识别出异常数据，但是消耗了大量的时间资源。詹静等人设计了一种新的可信 Modbus/TCP 通信协议[2]，提高了使用专用通信协议的 ICS 网络安全性，但是该协议的实现需要提供可信硬件模块，而且对协议进行认证也会影响通信的时间性能。GOLDENBERG N 等人提出了以寄存器值作为特征的异常检测模型[3]，以寄存器值是否处于正常值域范围来判断数据是否正常，但此模型的检测方法过于片面，忽略了 Modbus_TCP 协议通信过程中功能码字段的重要性。尚文利等人设计了一种粒子群优化算法(Partical Swarm Optimization，PSO)进行参数寻优的 PSO-SVM 算法[4]，通过功能码的频率识别 Modbus_TCP 的异常流量，但该方法只考虑功能码的作用，忽略了寄存器地址与功能码之间的对应关系。李超等人提出了单类支持向量机的算法[5]，该研究提取功能码和寄存器地址的组合对序列作为特征进行异常检测，分类效果显著，但是该方法数据预处理过程复杂，需要对数据集进行归一化处理才能进行模型训练，易造成时间资源的浪费。

综合前人研究成果的优点与不足，为了满足 ICS 的安全性需求，能够快速有效地发现 ICS 通信数据的异常状态以及系统潜在的威胁，本文对基于机器学习中的几种 Modbus_TCP 通信异常检测方法进行了对比研究，其中决策树算法相比于其他的机器学习方法有一定的优势：

(1)该算法对数据的要求程度不高，一般只需经过简单的预处理便可供模型使用，而不用像支持向量机和逻辑回归分类模型那样需要对数据进行归一化处理，才能获得较好的模型。

(2)该算法可以使用小样本进行模型训练，况且，即使是大样本，决策树分类模型也能在取得较好的分类效果的同时仅消耗较短的时间。而此优势又十分符合工业控制系统高实时性的要求。

本文通过实验仿真模拟工控系统中主、从站间的数据通信，并通过 Python 中的 Scapy 库设计异常数据对模拟的通信系统进行发包攻击，从而在异常检测终端获取正常行为和异常行为两种类型的原始数据，再将原始数据经过预处理后输入到决策树分类模型的算法中训练，通过实验测试能够有效地识别出 Modbus_TCP 流量中的正常数据和异常数据。

1 Modbus 协议与 Modbus_TCP 协议简介

Modbus 是一种位于应用层的通用传输协议[6]，常应用于工业控制系统之中。Modbus 协议报文由 4 部分构成：地址域、功能码、数据域和 CRC 校验域。其中，数据域包含的信息有：寄存器地址、需要操作的项目和域中数据的实际大小。

Modbus_TCP 协议是 Modbus 协议在 TCP/IP 协议上的扩展，由 3 个部分组成：Modbus 应用程序协议报头(Modbus Application Protocol，MBAP)、功能码和数据域。其中 MBAP 报头是 TCP/IP 协议用来识别 Modbus 应用数据的单元，包含事务单元标识符、协议标识符、长度和单元标识符四部分。两种模式的协议报文格式如图 1 所示。

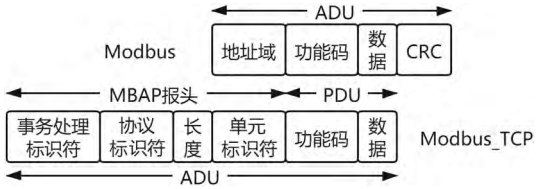

图 1 Modbus 协议和 Modbus_TCP 协议报文格式

由图 1 可见，Modbus 协议只定义了功能实现字段、循环冗余字段，Modbus_TCP 协议在 Modbus 协议的基础上增添了简单的识别字段，并没有设置任何可靠的防护机制，尤其是当系统接入到以 TCP/IP 协议为主的开放性网络中时，整个工业内网便完全暴露在外网环境里，这给了黑客可乘之机，他们可以轻而易举地获取工业内网的信息，甚至拦截、伪造数据重传，造成信息泄露或系统崩溃等恶劣影响。不过，Modbus_TCP 协议存在的缺陷在给工业控制系统带来巨大风险的同时，也提供了研究防护的契机。本文便是针对 Modbus_TCP 协议简捷性、开放性的特点来构造异常数据报文，模拟攻击环境，以实现异常检测。

2 基于决策树算法检测 Modbus_TCP 协议中的异常流量

2.1 决策树算法

决策树算法的核心思想是将测试数据集输入

到训练好的决策树分类模型中，从根节点开始，匹配数据集各列特征的相应属性，根据特征中不同的值寻找相应的分支，最终遍历到叶子节点，将叶子节点中存放的分类样本标签的值作为决策结果。本节中，通过信息增益[7]来选出最优特征以确定决策树的根节点及各子树的根节点。

信息熵作为衡量一个事物特征的无序程度的度量标准，能在数据集中选择出无序程度最大的那列特征作为决策树的划分节点。信息熵定义如下：

$$H=-\sum_{i=1}^{m} p_i \log p_i \tag{1}$$

其中 p_i 表示选择该分类的概率，m 为样本特征的列数。由式(1)可见，样本特征的不确定性越大(p_i 值越小)，信息熵 H 越就大。若将训练数据集定义为 X，且信息熵中的概率是由最大似然估计得到，则称信息熵为经验熵，用 $H(X)$ 表示。若用 $|X|$ 表示训练数据集 X 的样本个数，设训练集 X 有 n 个类 A_i，$i=1,2,\cdots,n$，$|A_i|$ 是 A_i 的样本个数。则：

$$H(X)=-\sum_{i=1}^{n} \frac{|A_i|}{|X|} \log \frac{|A_i|}{|X|} \tag{2}$$

选取最优特征需要计算各个特征的信息增益，而计算信息增益需引进条件熵的概念。条件熵表示在已知随机变量 Y 的情况下随机变量 X 的不确定性，用 $H(X|Y)$ 表示。

$$H(X|Y)=\sum_{i=1}^{n} p_i H(X|Y=y_i) \tag{3}$$

其中，$H(X|Y=y_i)$ 表示已知 $Y=y_i$ 的情况下 X 的条件分布概率的熵。

信息增益与条件熵密切相关，特征 Y 对训练集 X 的信息增益 $\mathrm{Gain}(X,Y)$ 定义为集合 X 的经验熵 $H(X)$ 与特征 Y 给定条件下 X 的条件熵 $H(X|Y)$ 之差，即：

$$\mathrm{Gain}(X,Y)=H(X)-H(X|Y) \tag{4}$$

2.2 决策树模型构建

(1)特征选取。在 Modbus_TCP 协议中，功能码是主站向从站发送控制信息的重要字段，倘若数据报文中的功能码出现异常，将导致执行器出现错误的操作。此外，寄存器地址与功能码有着很强的对应关系，它们之间的可靠性与系统的安全性同样息息相关，常用寄存器地址与功能码的对应关系如表 1 所示。Modbus_TCP 协议中其他报文字段的缺失或遭到非法篡改不会对系统造成恶劣影响，而且检测这些冗余的字段会占用很多存储资源和计算资源。因此，本文只选择功能码和寄存器地址作为决策树分类模型的输入特征，以满足异常检测的有效性和工业控制系统的实时性要求。

表 1 功能码与寄存器地址映射表

功能码	操作名称	寄存器地址范围
0x01/01	读线圈状态	00001~09999
0x02/02	读开关状态	10001~19999
0x03/03	读保持寄存器	40001~49999
0x04/04	读输入寄存器	30001~39999
0x05/05	写单个线圈	00001~09999
0x06/06	写单个保持寄存器	40001~49999
0x0F/15	写多个线圈	00001~09999
0x10/16	写多个保持寄存器	40001~49999

(2)数据的捕获与交叉验证。用抓包工具捕获主站与从站之间的通信流量，提取每条数据中的功能码和寄存器地址字段作为样本特征，并且对每一个样本标记好标签，构成原始数据集，将原始数据集按照 9:1 的比例进行交叉验证，生成训练集与测试集。

(3)决策树分类模型训练。该模型可由 Sklearn 框架进行训练，Sklearn 是机器学习中常用的 Python 第三方模块，里面封装了许多机器学习的方法，只需简单调用便可实现机器学习任务。训练步骤如下：

①提取训练数据集中的分类样本标签，用式(2)计算其信息熵。

②分割功能码和寄存器地址这两列特征，提取训练数据集中的功能码，统计功能码值相同的样本数量并计算出其在总样本数量中的概率，然后再用式(2)计算不同功能码值的信息熵，最后用式(3)计算出整列功能码的条件熵；同理计算出寄存器地址的条件熵。

③由式(4)可见，信息增益的值是由信息熵和条件熵确定的，条件熵越大信息增益就越小。用式(4)可计算出功能码和寄存器地址这两列特征的信息增益，比较两者信息增益的大小。

④取信息增益较大的特征作为最优特征并成为决策树根节点，次优特征则成为决策树的中间节点，将最后的二分类决策结果作为决策树的叶子节点。由此，得出一棵高度为 3 的决策树训练模型。

(4)测试模型结果。调节 Sklearn 模块中相应的参数，将测试数据集中的功能码和寄存器地址这两列特征输入到决策树训练模型，从决策树的根节点

开始，匹配数据集中最优特征的属性，并根据其值寻找相应分支进入下一节点，继而匹配数据集中次优特征的属性，再根据其值寻找相应的分支得到叶子节点的值，正常数据的叶子节点的值为 1，反之，叶子节点的值为 0。

决策树模型构建流程图如图 2 所示。

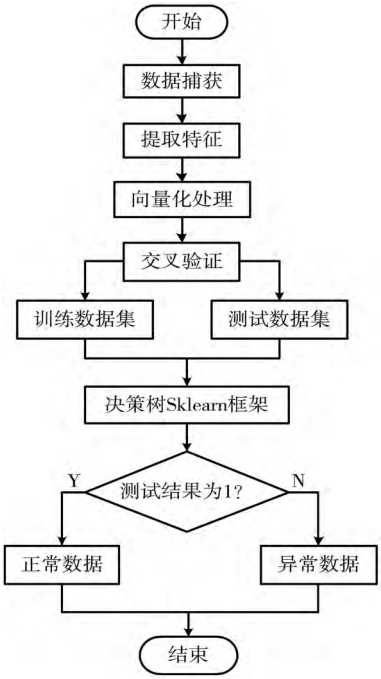

图 2 决策树算法分类模型构建流程图

3 实验仿真

3.1 实验环境搭建

本次实验通过 Windows 7 系统进行仿真，使用 Modbus Poll 工具与 Modbus Slave 工具分别模拟工业控制系统中操作主站与操作从站。Modbus Poll 是一个非常强大的 Modbus 开发调试工具，支持进行主站设备仿真，并支持多文档接口，可以同时监控和调试多个从站设备。Modbus Slave 是从站设备仿真软件，主要功能是接收主站设备的命令并发出响应，可用于测试主站设备，观察 Modbus 通信中的各种报文。Modbus Poll 和 Modbus Slave 有着相同的操作界面，它们支持功能码 01、02、03、04、05、06、15、16、22 和 23，支持 ASCII、TCP/IP 和 ModbusRTU 等协议。Scapy[8]是 Python 中一个强大的交互式数据包操作程序，主要有路由跟踪、扫描测试、发包攻击和网络探测与发现等功能。本实验建立起 Modbus Poll 与 Modbus Slave 模拟主站与从站之间的正常通信，通过使用 Modbus_TCP 协议产生了正常的数据报文，Scapy 攻击源接入 Modbus Poll 与 Modbus Slave 的通信链路之间，不断伪造 Modbus_TCP 数据报文，发往系统。为有效处理模拟通信中的异常数据报文，在网关处接入异常检测终端，使用 Python3.7 集成环境，安装与决策树相关的机器学习方法和 Sklearn 模块来分析处理数据。具体模拟通信环境如图 3 所示。

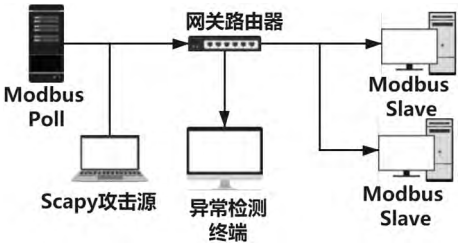

图 3 模拟通信拓扑图

3.2 采集数据

异常检测终端处安装 Wireshark 抓包工具，使用其自带的命令行工具 Tshark 捕获数据。Tshark 是 Wireshark 的一个命令行工具，可以通过调用命令对 Modbus_TCP 协议网络数据包中功能码和寄存器地址等重要字段进行提取，从而实现对数据的处理。

首先，打开 Tshark 抓包工具，设置相应的应用参数，使 Tshark 只捕获其中端口号为 502 的 Modbus_TCP 数据报文；然后，调用相应的命令提取每条数据中的功能码字段和寄存器地址字段。

将提取到的功能码字段和寄存器地址字段分别作为输入数据集的两列特征，并且对每一个样本设置好标签，预处理成决策树算法能够处理的数据集。本实验将预处理好的数据集分别随机抽取 160、320、480 个样本的 3 组数据作为原始数据集。随后进行交叉验证，将原始数据集一分为二，随机提取原始数据集的 10%作为测试集，剩下的 90%作为训练集，用测试集验证训练模型；3 组不同样本数量的数据重复进行交叉验证，提取测试集与训练集各 10 对。

3.3 决策树分类模型分类结果与分析

调用 Sklearn 模块进行模型训练。将预处理好的 10 对训练集和测试集分别输入到决策树分类算法模型之中，通过 Sklearn 模块中的预测语句，得出决策树分类模型的准确率。3 组样本不同的数据集测试出模型的准确率如图 4 所示。

由图 4 可见，本实验的决策树分类模型的准确率会随样本数的增加而逐渐增加，最终达到接近于

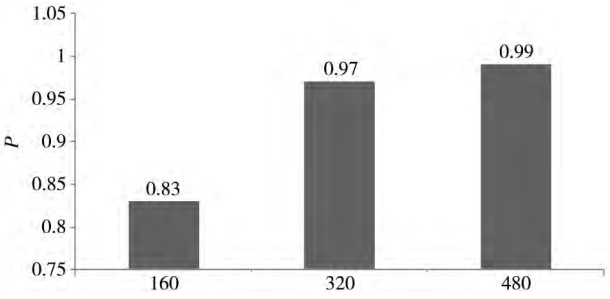

图 4 不同样本数量的决策树分类模型准确率柱状图

100%的稳定状态。而且在接近于 100%准确率所使用的样本数量并不高。总而言之，决策树是一个具有支持小样本训练且分类精度高的分类模型。

4 模型对比评估

本文设置了朴素贝叶斯分类模型[9]、逻辑回归分类模型[10]和传统支持向量机分类模型[11]作为对比实验以验证决策树分类模型的可行性与优越性。在设置逻辑回归和传统支持向量机的实验时，需要将所有实验的数据集转成二进制模式，进行归一化处理，才能进行模型训练。

本实验重点对 4 个模型分离出正常流量和异常流量的准确率、误报率和漏报率进行评判。此外，工业控制系统对时间性能的要求非常高，高准确率、高延迟的分类模型对于系统而言并没有实际的研究意义，因此，本实验也将 4 个分类模型的算法运行时间作为评判模型优劣的一个重要指标。

4.1 准确率、误报率和漏报率对比

实验对 4 个分类模型分别取 160、320、480 个样本的 3 组数据集分别进行 10 次分类测试，调节 Sklearn 模块参数进行训练与测试，并生成各自对应的准确率、误报率和漏报率，分别对每组 10 次测试的结果取平均值。决策树等 4 个分类模型的分类准确率对比分析图如图 5 所示。

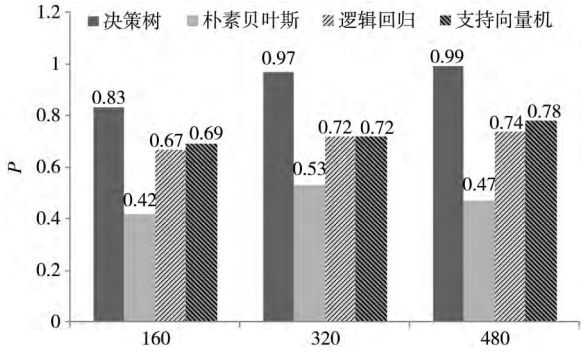

图 5 决策树等 4 个分类模型算法准确率柱状图

由图 5 可见，决策树分类模型的准确率远远高于其他 3 个分类模型的准确率。

此外，误报率与漏报率是在准确率的基础上评价模型优劣的指标，误报率太高将导致系统频频拉响异常警报信号，影响正常通信；漏报率太高则说明模型无法对异常的流量数据进行识别，将导致系统无法拉响异常警报信号而陷入严重的危险之中。决策树等 4 个分类模型误报率和漏报率对比分析图如图 6 所示，图中决策树的误报率和漏报率是最低的，模型最优；朴素贝叶斯的这两项指标在 0.5 处保持平衡；逻辑回归和传统支持向量机则呈现高误报率、低漏报率。功能码字段和寄存器地址字段所组合成的正常数据和异常数据的点在直角坐标系上分布较混乱，复杂度较高，因此，朴素贝叶斯分类模型对其分类困难；而逻辑回归分类模型和传统支持向量机分类模型对数据正负样本的平衡性要求较高。总而言之，在面对分布混乱复杂，正负样本不平衡的数据时，决策树分类模型通过选择分支能更容易地实现分类，从而体现出一定的优势。

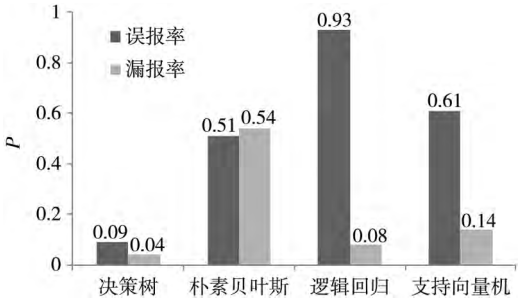

图 6 决策树等 4 个分类模型算法误报率和漏报率柱状图

4.2 时间性能对比

对 4 个分类模型分别进行 10 次测试，计算 4 个分类模型算法每次测试运行的时间，对其取平均值，对比分析结果如图 7 所示。

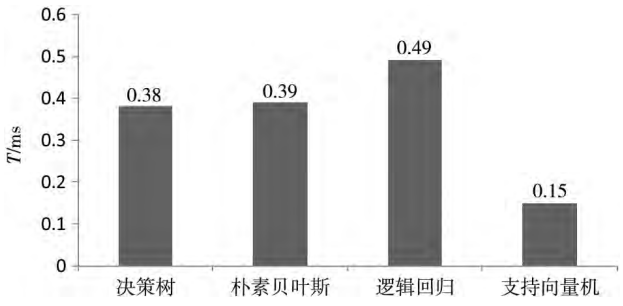

图 7 决策树等 4 个分类模型算法运行时间柱状图

图 7 中的时间表示平均每处理一个样本，分类模型的算法所耗费的时间。在 4 个分类模型的算法中，传统支持向量机分类模型的算法运行时间相对最短，决策树算法次之。但是对于工业控制系统的高实时性特点，毫秒级的处理时间显然是符合要求的。而且，决策树在数据预处理阶段无需作归一化操作，相比于逻辑回归和传统支持向量机节省了大量的时间资源，综上所述，决策树分类模型在时间性能方面更优。

5 结论

针对工业控制系统中的 Modbus 协议存在的诸多安全问题，本文对基于机器学习的几种 Modbus_TCP 通信异常检测方法进行对比研究。对决策树、朴素贝叶斯、逻辑回归和支持向量机 4 个分类模型进行实验评估，结果表明决策树分类模型在满足工业控制系统的高可用性、高可靠性、高实时性需求的基础上，能够检测出 Modbus_TCP 协议中的网络通信是否存在异常，分类准确率较高，消耗时间较短，体现出了一定的优越性。

参考文献

[1] EREZ N, WOOL A. Control variable classification, modeling and anomaly detection in Modbus/TCP SCADA systems[J]. International Journal of Critical Infrastructure Protection, 2015, 10(C): 59-70.

[2] 詹静，杨静. 基于远程证明的可信 Modbus/TCP 协议研究[J]. 工程科学与技术，2017，49(1)：197-205.

[3] GOLDENBERG N, WOOL A. Accurate modeling of Modbus/TCP for intrusion detection in SCADA systems[J]. International Journal of Critical Infrastructure Protection, 2013, 6(2): 63-75.

[4] 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[J]. 电子学报，2014，42(11)：2314-2320.

[5] 李超，蔡宇晴，贾凡，等. 工业控制系统中基于单类支持向量机异常检测方法研究[J]. 微型机与应用，2017，36(23)：9-12.

[6] 罗旋，李永忠. Modbus TCP 的安全机制研究与实现[J]. 信息技术，2019，43(1)：15-19.

[7] 许鸿坡，陈伟. 基于决策树的远程控制协议字典攻击检测[J]. 计算机技术与发展，2019，29(6)：105-111.

[8] 李兆斌，茅方毅，王瑶君，等. Scapy 在网络设备安全性测试中的应用[J]. 北京电子科技学院学报，2016，24(4)：73-77.

[9] 李腾飞. 基于多项式朴素贝叶斯算法的垃圾邮件过滤器的设计与实现[J]. 科技资讯，2018，16(33)：1-2.

[10] 侯爱华，高伟，汪霖. 基于逻辑回归模型的流量异常检测方法研究[J]. 工程数学学报，2017，34(5)：479-489.

[11] 高超，许翰林. 基于支持向量机的不均衡文本分类方法[J]. 现代电子技术，2018，41(15)：183-186.

(收稿日期：2020-06-05)

作者简介：

陈鑫龙(1995-)，男，硕士研究生，主要研究方向：工业控制系统与信息安全。

陈志翔(1982-)，男，博士，副教授，主要研究方向：工业控制系统与信息安全。

周小方(1963-)，通信作者，男，硕士，教授，主要研究方向：嵌入式系统。E-mail：2334226013@qq.com。