文章编号:1671-4598(2021)05-0174-05 DOI:10.16526/j.cnki.11—4762/tp.2021.05.035 中图分类号:TP301.6 文献标识码:A

智能化网络安全防攻击检测中数据抽取和分析

杨宗跃
(国网思极网安科技(北京)有限公司,北京 102211)

摘要:针对传统智能化网络安全防攻击检测平台处理数据效率低、误差大等问题,本研究提出一种新型的解决方案;该方案数据抽取模型和大数据分析构建智能化网络安全防攻击检测平台,采用特征模板、卷积神经网络算法模型和条件随机场算法3种方法结合构建出数据抽取模型来抽取网络安全检测数据;其中,利用特征模板提取局部特征向量并进行语句转换得到初始局部向量序列,通过CNN算法对每个网络安全检测数据样本进行卷积和聚合,并提取其特征信息,将语义特征和局部特征相结合经过条件随机场算法进行序列标记,并抽取最优的特征向量序列,最后通过置信传播改进的逻辑回归模型进行分析;实验表明,本研究所提方案克服了现有技术存在的不足,显著提高了处理数据效率和精准度,在数据量为2GB的环境下,经过对数最大似然损失函数得出的损失值只有0.35。

关键词:智能网络;防攻击检测;数据抽取模型;逻辑回归;置信传播

Data Extraction and Analysis in Intelligent Network Security Anti—attack Detection

Yang Zongyue
(State Grid Network Security (Beijing) Technology Co., Ltd., Beijing 102211, China)

Abstract: Aiming at the problems of low data processing efficiency and large error of traditional intelligent network security attack detection platform, this paper proposes a new solution. In this scheme, the data extraction model and big data analysis are used to build an intelligent network security attack detection platform. The network security inspection data extraction model is constructed by combining feature template, convolution neural network algorithm model and conditional random field algorithm to extract network security detection data. Among them, the local feature vector is extracted by feature template, and the initial local vector sequence is obtained by sentence conversion. Each network security monitoring data sample is convoluted and aggregated by CNN algorithm, and its feature information is extracted. The semantic feature and local feature are combined, and the sequence is marked by conditional random field algorithm, and the optimal feature vector sequence is extracted The improved logistic regression model of belief propagation was analyzed. Experiments show that the proposed scheme overcomes the shortcomings of existing technologies, and significantly improves the efficiency and accuracy of data processing. In the environment of 2GB data, the loss value obtained by the logarithmic maximum likelihood loss function is only 0.35.

Keywords: Intelligent network; anti attack detection; data extraction model; logistic regression; belief propagation

0 引言

在信息智能化不断发展的时代,许多国家接连发生了大型网络攻击事件,各大型企业产业经济遭到史无前例的重创。大量案例表明,智能网络时代给企业的安全带来了全新的挑战[1]。经过实例分析,黑客对物联网等重要设施的攻击,通常都是从终端发起,攻击类型复杂、终端防护受自身条件和运行环境的限制,以及复杂多源的数据类型为后续数据处理给网络安全防范工作带来了极大的困难[2]。因此,如何提高网络安全风险防范效率,减少数据处理的时间开销,提高处理速度,是后续研究工作中需要解决的主要问题[3]。

针对上述存在的问题,许多学者发表了自己研究的技术方案。文献[4]公开了一种网络流量数据抽样技术,虽然在一定程度上提高了数据处理效率,但是只能处理高频率的流量数据,而忽略了低频率的流量数据,存在处理不平衡问题。文献[5]提出基于交叉验证优化贝叶斯分类法,对网络安全检测数据进行有效地分类处理,但是随着交叉数目的不断增加,数据预处理过程耗时会逐渐增加,导致效率大打折扣。

1 总体方案设计

针对上述技术存在的不足,本研究设计出新型的智能化网络安全防攻击检测平台,全面分析网络风险因素,以提高对网络风险因素的感知、预测和防范能力。关于网络安全防攻击检测平台总体框架图如图1所示。

从图1可以看出,网络安全检测平台主要是通过物联网和企业的业务系统中获取数据,利用网络采集探针在关

收稿日期:2020-10-08; **修回日期**:2020-10-26。

作者简介:杨宗跃(1983-),男,北京人,学士,网络安全工程师,主要从事电网信息安全、数据安全方向的研究。

引用格式:杨宗跃.智能化网络安全防攻击检测中数据抽取和分析[J].计算机测量与控制,2021,29(5):174-178.

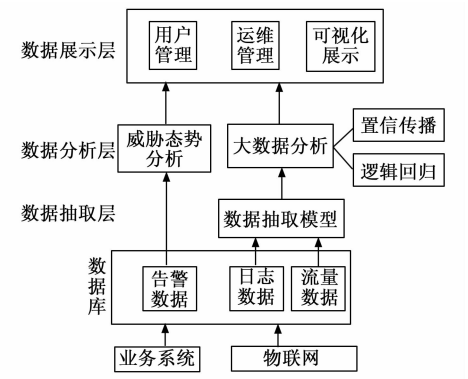

图 1　网络安全防攻击检测平台总体框架图

键网络节点进行实时检测，并将采集得到的不同结构的数据进行合理的存储，为数据抽取模型的特征提取提供足够的样本信息。在分析层中进行合理的数据分析，并将分析结果传达至上层管理，根据决策者、管理人员和运维人员不同的需求和关注重点，通过可视化分析技术，挖掘出恶意软件或流量数据隐藏的数据信息，并进行多种态势的多维度展示，并且支持预警通告和应急处置[6-7]。

2　关键技术

为了解决复杂多源的网络安全检测数据处理工作复杂的问题，本研究构建数据抽取模型和数据分析模型两种技术来解决该问题，下面将分别阐述。

2.1　网络安全数据抽取模型的构建

由于网络安全检测数据类型多样，且不断地会有例如恶意软件、漏洞以及补丁等新的数据出现，因此基于分词的方法识别率较低。针对上述问题，本研究基于卷积神经网络（CNN）模型结合特征模板构建出一种新型的网络安全数据抽取模型，如图 2 所示。

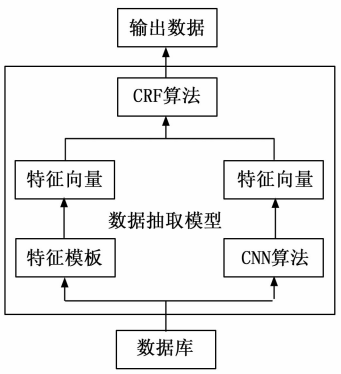

图 2　网络安全检测数据抽取模型

整个构建过程首先要根据网络安全防攻击检测数据库手工生成少量特征模板，并提取局部特征向量，然后根据把网络安全检测数据特征向量进行语句转换得到初始局部向量序列；其次，通过 CNN 算法对每个网络安全检测数据样本进行卷积和聚合，并提取其特征信息；最后，将语义特征和局部特征相结合经过条件随机场（CRF）算法进行序列标记，并抽取最优的特征向量序列。下面将分别阐述该过程中各部分的具体内容。

2.1.1　特征模板

特征模板是根据选取大量的数据特征并建立特定的模板，在数据抽取模型中便于之后识别数据的特征提取。特征模板的建立取决于“模板窗口”，窗口过大则会出现过拟合现象，窗口过小则提取特征向量十分有限，造成网络安全检测数据模型的识别效率较低，因此设计特征模板的模板窗口大小要十分合理。关于特征模板的构建过程如下：

首先设网络安全检测数据特征一系列为 $w\,[-1,\,0]$，$w\,[0,\,0]$，$w\,[1,\,0]$，…，$w\,[i,\,j]$。其中，w 表示网络安全检测数据信息字符，括号内的第一个数字 i 表示相对 w 的位置，第二个数字 j 表示特征列数。通过对特征函数 f 定义为：

$$f_k(y,w)=\sum_{x=1}^{n} f_k(y_{x-1},y_x,w,x) \tag{1}$$

其中：y 表示当前标记的网络安全检测数据信息字段；k 表示特征函数数目，k 通常取自然数；x 表示当前字段位置。在通常情况下，特征函数 f 得出的数值是二值函数，式（1）的含义是取在 x 位置每个特征函数的总和[8-9]。设赋予特征函数的权重向量 $\boldsymbol{Z}$ 为：

$$\boldsymbol{Z}=(z_1,z_2,\cdots,z_k)^T \tag{2}$$

之后，将所有当前标记的网络安全检测数据特征信息转换为特征向量，得到：

$$\boldsymbol{F}(y,w)=(f_1(y,w),f_2(y,w),\cdots,f_k(y,w))^T \tag{3}$$

其中：$\boldsymbol{F}$ 表示所有特征向量的总序列。

2.1.2　CNN 算法

为了有效的提取网络安全检测数据字符级特征，本研究采用 CNN 算法模型处理那些细粒度高的字符特征。CNN 算法能够自适应地从具体到抽象地特征信息，并且可以拥有不同结构的神经网络框架，灵活性很高[10-11]。关于基于 CNN 的特征提取流程图如图 3 所示。

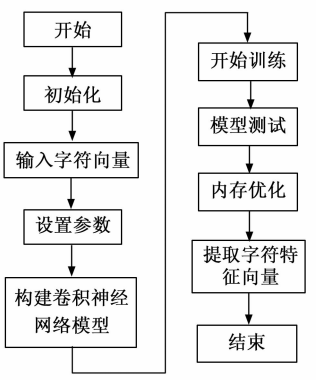

图 3　基于 CNN 的字符特征提取流程图

如图 3 所示，在输入网络安全检测数据字符向量后，要先设置相关参数、损失函数和优化器。相关参数依然是由迭代次数、批处理以及学习速率组成，为了减少内存消耗，通过添加神经网络压缩加速技术对内存进行优化。通

过将模型测试和优化交替判断处理，可以使训练时间更快[12-13]。在构建好神经网络模型之后，将分类交叉熵函数作为损失函数进行模型测试。

在整个特征提取过程中构建卷积神经网络模型是最重要的步骤，其主要由卷积层、池化层、全连接层和输出层四部分组成：

卷积层对于网络安全检测数据来说相当于一种滤波器，与滤波器所不同的是卷积是通过卷积核的不同对输入进行训练处理，提高了效率，极大地减少了参数量；池化是利用卷积核来减少数据的参数个数并依然能进行特征提取的过程。池化操作虽然丢失了一些信息，但保持了网络安全检测数据的平移和扩展的不变性；全连接层就是将每一层的神经元都要与下一层所有神经元相连，也是为了将池化后的网络安全检测数据特征信息进行学习权重系数并分类[14-16]。本研究通过固定输入网络安全检测数据大小以及全连接层系数矩阵，为输出层提供更加突出的特征信息。

2.1.3 CRF 算法

由于存在不能独立的抽取网络安全检测数据特征向量的问题，因此本探究通过链式 CRF 算法计算整体上特征向量标签序列的概率并得出损失值。首先，输入网络安全检测数据标签特征序列 Y 为：

$$X=(x_1,x_2,\cdots x_n) \tag{4}$$

$$Y=(y_1,y_2,\cdots y_n) \tag{5}$$

其中：X 表示网络安全检测数据特征序列，Y 是 X 的标签序列，括号中每个字母代表着一个特征向量。之后计算每个输入特征向量在 t 时刻的标签权重 M：

$$\begin{cases} M_1=Z\cdot F(y,w) \\ M_2=W_cO_t \\ M=M_1+M_2 \end{cases} \tag{6}$$

其中：M_1 和 M_2 分别表示经过特征模板和 CNN 算法得出的权重值；Z 和 F 分别表示权重向量和特征向量的总序列；W_c 和 O_c 分别表示 CNN 算法中的权重矩阵和输出层输出结果[17]。

计算在输入序列 X 的情况下产生标签序列 Y 的概率 P 的表达式为：

$$P=\frac{e^{M(x,y)}}{\sum_{y\in Y}M(x,y)} \tag{7}$$

本研究采用对数最大似然来表示损失函数，最终得到：

$$\log P=M(x,y)-\log\left(\sum_{y\in Y}e^{M(x,y)}\right) \tag{8}$$

根据公式（8）损失函数值输出最优标签的网络安全检测数据，得到结果 H 为：

$$H=\operatorname{argmax}M(x,y) \tag{9}$$

2.2 基于逻辑回归的网络安全分析方法

在网络安全防攻击检测中，通过在数据抽取模型中得到网络安全的特征向量，本研究采用逻辑回归对攻击检测中的网络安全数据进行分析，应用二元分类解决数据难处理问题[18-19]。本研究中网络安全数据分析通过逻辑回归模型来实现，下面说明具体构建过程：

设网络安全检测数据特征序列为 $A=[a_1,a_2,\cdots,a_n]^T$，与其相对应的类集合 $B=[b_1,b_2,\cdots,b_n]$，设 C 是两个预定类集合，通过 Logit 函数将特征向量 a 映射到两个预定类集合中的某一个得到二分类的逻辑回归模型[20]：

$$P(b=c\mid a)=\frac{1}{1+e^{\lambda}} \tag{10}$$

$$\lambda=k_0+k_1a_1+k_2a_2+\cdots k_na_n \tag{11}$$

其中：k 为网络安全检测数据特征序列的权重系数，P 表示特征向量 a 的攻击概率。

由于在企业网络安全防攻击检测中对不同恶意软件分析的要求，因此逻辑回归模型得到具有概率意义的结果将更好。逻辑回归模型在处理每个网络安全检测数据样本二元分类结果，对应出的一个处于 0～1 之间的概率值 P，可以表明分类结果的置信度，即概率 P 可以作为发生网络安全风险可能大小的衡量标准[21]。但在实际应用中，网络安全检测数据样本二元分类的结果会存在一定的偏差，即样本多数类和少数类的问题。

针对这种问题，本研究采用置信传播技术对逻辑回归模型进行改进，增添了网络安全防攻击独立特征条件概率，每当新的数据样本被抽取时便可快速提取特征用于逻辑回归模型进行分析，进一步获取攻击的置信度。而在算法模型改进的层面上，通常的方法是使算法在不平衡分类问题上表现更好。其主要方法是通过算法对决策面进行修正，使其偏向少数类，从而提高少数类的识别率，比如贝叶斯网络模型改进的概率密度算法。

置信传播通常是计算置信度与真实值比较来判断网络安全检测数据是否判定是否处于网络安全告警状态。关于逻辑回归模型中置信传播过程如图 4 所示。

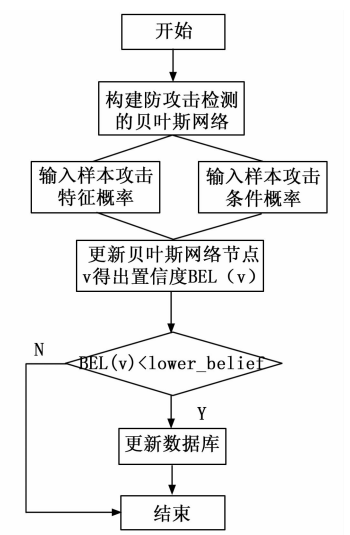

图 4 置信传播流程图

在网络安全防攻击检测过程中，条件概率由状态可以分为攻击属性和良性属性。关于置信度 $BEL(v)$ 的计算中需要贝叶斯网络中每个节点的条件概率表，其中覆盖了每

个可能状态的边界概率 P_2。通常情况下条件概率表中攻击属性和良性属性的条件概率相同，这也是理想情况下网络安全检测数据样本二元平衡分类的结果。在攻击属性下的置信度 $BEL(v)$ 的计算过程为：

$$BEL(v) = P_2^1 \cdot P_2^2 \cdots \cdot P_2^n \tag{12}$$

式（12）表示每个网络安全检测数据样本攻击条件概率之积即为攻击节点的置信度，良性属性下同理。

综上所述，整个网络安全分析方法先从数据抽取模型中得到所有数据样本，优点在于传递给置信传播模型前即可确定未知样本的恶意攻击概率。经过逻辑回归模型分析出该样本特征为攻击属性或良性属性，大幅度提高了效率。置信传播过程中使用恶意攻击的所有条件概率的特征属性计算置信度 $BEL(v)$，这种方法能够使输出的结果不受独立条件概率的影响，提高对样本分析的精准度[22-23]。

3　实验与分析

为了验证本研究网络安全防攻击检测平台的可靠性和实用性，下面将进行实验。

3.1　实验环境与数据样本

关于实验环境可分为硬件环境和软件环境，其中硬件环境为机台为 CentOS6.8（x64）操作系统，Intel（R）Xeon（R）CPU E5－2640 v2、2.00 GHz 主频、千兆网卡、8 核 16 G 内存、512 GB 硬盘。软件的操作系统为 Windows10，JDK5.0。

关于实验设置本研究采用一主机 6 个服务器节点来构建网络安全防攻击检测系统，在服务器节点网络流量数据传输末端设置用户服务器，评估用户空间恶意软件和内核级 Rootkit 攻击能力。关于网络安全防攻击检测过程中的恶意软件部分类型如表 1 所示。

表 1　网络安全防攻击检测过程中的恶意软件

恶意软件	执行空间	特性
反向恶意代码	用户空间	建立外反向连接
C&C 僵尸网络	用户空间	建立主从僵尸网络连接
Xing Guo Quan	内核空间	在内核执行且建立外连接
Azazel	内核空间	在内核执行且建立外连接

3.2　实验内容与结果分析

在上述的模拟仿真实验中，下面对本研究的系统进行验证，将带有本研究置信传播（BP）模型和未带有 BP 模型的逻辑回归分析方法在网络安全检测中进行数据分析，采用的数据从表 1 中的网络安全检测恶意软件数据随机选取一种，评估攻击概率与真实值在 0～2 GB 数据量下的对比，通过 MATLAB 软件进行仿真，得出曲线图如图 5 所示。

从图 5 可以看出，在不同网络安全检测数据量环境下，采用 BP 模型得到的攻击概率 P 与真实值相差较小且趋于稳定，不采用 BP 模型得到的攻击概率 P 与真实值相差较大且波动幅度明显。从该实验结果表明本研究采用置信传播技术改进逻辑回归模型能有效提高网络安全防攻击检测数据分析的精准度。

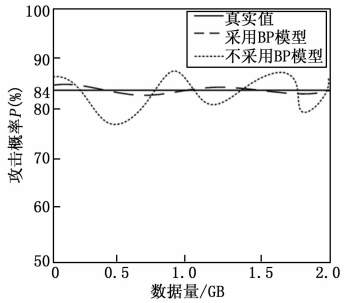

图 5　对比曲线图

为了验证本研究所设计的数据抽取模型的优势，本研究以文献［5］中基于交叉验证优化贝叶斯分类法作为对比，采用不同方法计算 0～2TB 网络安全检测数据量范围内损失值，通过 MATLAB 软件系统进行仿真对比，对比结果图如图 6 所示。

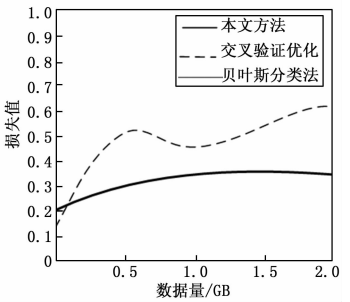

图 6　损失值对比结果图

从图 6 可以看出，本研究所采用的数据抽取模型方法比交叉验证优化贝叶斯分类法的损失值更低，网络安全检测数据抽取性更加高效，这充分表明本研究的数据抽取模型更加适用。

4　结束语

本研究设计出新型的智能化网络安全防攻击检测平台，构建数据抽取模型提高网络安全检测数据特征抽取的精准度。通过分析存在的潜在威胁和恶意软件，评估网络受到攻击的概率，最后通过实验验证了本研究的网络安全防攻击检测平台的适用性和可靠性。实验结果表明，本研究的数据抽取模型能够产生最优估计值，而采用基于置信传播改进逻辑回归模型处理数据更接近真实值。随着技术的不断发展，对于智能化网络安全检测平台采集精准度和全面性要求会更高，本研究仍旧存在诸多不足，有待进一步的研究。

参考文献：

［1］赵悦品．网络信息安全防范与 Web 数据挖掘系统的设计与实现［J］．现代电子技术，2017，40（4）：61－65.

［2］葛元鹏，蔡晓明，周　晟，等．电网企业信息安全防护体系建设［J］．电子科技，2015，28（7）：186－188.

［3］Nikiforakis N，Kapravelos A，Joosen W，et al. On the workings and current practices of web－based device fingerprinting［J］.

IEEE Security & Privacy, 2014, 12 (3): 28 - 36.
[4] 翟金凤, 孙立博, 鲁 凯, 等. 基于 Counting Bloom Filter 的流抽样算法研究 [J]. 计算机工程, 2018, 44 (8): 279 - 284.
[5] 侯佳音, 史淳樵. 网络信息安全问题研究及防护策略设计与研究 [J]. 电子设计工程, 2015, 23 (22): 158 - 160.
[6] 王 惠, 刘 霓, 刘东全. 政务部门网络安全态势感知系统构建研究 [J]. 中国信息安全, 2019, 111 (3): 80 - 81.
[7] 杨 安, 孙利民, 王小山, 等. 工业控制系统入侵检测技术综述 [J]. 计算机研究与发展, 2016, 53 (9): 2039 - 2054.
[8] 石 军. 信息安全隐患展示系统的研究与开发 [J]. 现代电子技术, 2017, 40 (8): 11 - 13.
[9] 赵 刚, 宫义山, 王大力. 考虑成本与要素关系的信息安全风险分析模型 [J]. 沈阳工业大学学报, 2015, 37 (1): 69 - 74.
[10] 王 锴, 李志华, 黄 凡, 等. HyperSpector: 基于 UEFI 的 VMM 动态可信监控基的设计与实现 [J]. 网络与信息安全学报, 2016, 2 (12): 47 - 55.
[11] 陈亚亮, 戴沁芸, 吴海燕, 等. Mirai 僵尸网络恶意程序分析和监测数据研究 [J]. 网络与信息安全学报, 2017, 3 (8): 35 - 43.
[12] 何 远, 张玉清, 张光华. 基于黑盒遗传算法的 Android 驱动漏洞挖掘 [J]. 计算机学报, 2017, 40 (5): 1031 - 1043.
[13] 邓渊浩. 基于硬件虚拟化的内核漏洞监测系统的设计和实现 [D]. 镇江: 江苏大学, 2015.
[14] 付 钰, 李洪成, 吴晓平, 等. 基于大数据分析的 APT 攻击检测研究综述 [J]. 通信学报, 2015, 36 (11): 1 - 14.
[15] 刘 威, 刘 尚, 白润才, 等. 动态数据约简的神经网络分类器训练方法研究 [J]. 智能系统学报, 2017, 12 (2): 258 - 265.
[16] Friedberg I, Skopik F, Fiedler R. Cyber situational awareness through network anomaly detection: state of the art and new approaches [J]. Elektrotechnik und Informationstechnik, 2015, 132 (2): 101 - 105.
[17] 陆万青. 网络信息安全对攻击风险预测仿真 [J]. 计算机仿真, 2017 (11): 316 - 319.
[18] 孙 宇. 基于抽样流的网络流量异常检测技术研究 [D]. 北京: 北京交通大学, 2018.
[19] 王丽娜, 谈 诚, 余荣威, 等. 针对数据泄露行为的恶意软件检测 [J]. 计算机研究与发展, 2017, 54 (7): 1537 - 1548.
[20] 杨宏宇, 唐瑞文. 基于电量消耗的 Android 平台恶意软件检测 [J]. 清华大学学报, 2017, 57 (1): 44 - 49.
[21] 金志刚, 苏 菲. 基于 FSVM 与多类逻辑回归的两级入侵检测模型 [J]. 南开大学学报 (自然科学版), 2018, 51 (3): 1 - 6.
[22] 赵真灵. 基于置信传播的大规模 DAS 系统检测和预编码算法理论分析与应用 [D]. 南京: 东南大学, 2016.
[23] Verma S, Kawamoto Y, Fadlullah Z M, et al. A survey on network methodologies for real－time analytics of massive IoT data and open research issues [J]. IEEE Communications Surveys & Tutorials, 2017, 19 (3): 1457 - 1477.

(上接第 173 页)

[3] Ibe E, Taniguchi H, Yahagi Y, et al. Impact of Scaling on Neutron－Induced Soft Error in SRAMs From a 250 nm to a 22 nm Design Rule [J]. IEEE Transactions on Electron Devices, 2010, 57 (7): 1527 - 1538.
[4] Tipton A D, Pellish J A, Reed R A, et al. Multiple－bit upset in 130nm CMOS technology [J]. IEEE Transaction on Nuclear Science, 2006, 53 (6): 3259 - 3264.
[5] Swift G M, Guertin S M. In－flight observation of multiple－bit in DRAMs [J]. IEEE Transactions on Nuclear Science, 2001, 47 (6): 2386 - 2391.
[6] Sang P P, Lee D, Roy K. Soft－Error－Resilient FPGAs Using Built－In 2－D Hamming Product Code [J]. IEEE Transactions on Very Large Scale Integration Systems, 2012, 20 (2): 248 - 256.
[7] Shirvani P P, Sexena N R, McCluskey E J. Software－implemented EDAC protection against SEUs [J]. IEEE Transactions on Reliability, 2000, 49 (3): 273 - 284.
[8] Roger C. Goerl, Paulo R. C. Villa, Letícia B. Poehls, et al. An efficient EDAC approach for handling multiple bit upsets in memory array [J]. Microelectronics Reliability, 2018, 88 - 90.
[9] 贺兴华, 卢焕章, 肖山竹, 等. 基于改进型 (14, 8) 循环码的 SRAM 型存储器多位翻转容错技术研究 [J]. 宇航学报, 2010, 31 (3): 803 - 810.
[10] 李晓花, 王雅云, 于 锋, 等. 星载计算机 SRAM 抗单粒子多位翻转技术 [J]. 计算机工程与设计, 2015, 36 (6): 1519 - 1523.
[11] 刘小汇, 张 鑫, 陈华明. 基于一种交织码的多位翻转容错技术研究 [J]. 信号处理, 2012, 28 (7): 1014 - 1020.
[12] Ben Cooke. ReedMuller Error Correcting Codes [J]. Mit Undergraduate Journal of Mathematics, 1999, 21 - 26.
[13] Shu L, Daniel J, Costello Jr. 差错控制编码 (原书第 2 版) [M]. 晏 坚, 何亚元, 潘亚汉, 等译. 北京: 北京机械工业出版社, 2007.
[14] Eiji O, Jing Zhigang, Roberto R C, et al. Concurrent round－robin－based dispatching schemes for Clos－network switches [J]. IEEE / ACM Transactions on Networking, 2002, 10 (6): 830 - 844.
[15] Xilinx. XAPP715: Multiple bit error correction [Z]. 2004, 11.
[16] 梁 健, 张润宁, 赵 帅. 一种针对 COTS 器件的抗辐射加固方法 [J]. 航天器工程, 2016, 25 (4): 81 - 86.
[17] 郝亚男, 高 欣, 许仕龙. SRAM 型 FPGA 的 SEU 容错技术研究 [J]. 中国集成电路, 2015, 24 (10): 31 - 36.
[18] Liew S Y, Lee T T. Bandwidth assignment with QoS guarantee in a class of scalable ATM switches [J]. IEEE Transactions on Communications, 2000, 48 (3): 377 - 380.
[19] 冯水春, 孟 新, 毛博年, 等. 基于 (16, 8) 准循环码的星载 FPGA 有限状态机容错设计 [J]. 北京邮电大学学报, 2014, 37 (1): 85 - 89.