

一种基于日志的攻击模式挖掘方法

龙 门¹, 夏靖波¹, 张子阳²

(1 空军工程大学 电讯工程学院, 陕西 西安 710077; 2 空军工程大学 工程学院, 陕西 西安 710038)

摘 要: 建立“择优”和“劣汰”过程结合的攻击模式挖掘模型对日志信息进行分析。“择优”过程运用改进正交化算法的稀疏化策略对数据集进行重新组合,提高了最小二乘支持向量机分类模型的效率。“劣汰”过程提出了基于相关度检测法的改进滑动窗策略,剔除了虚假攻击事件,避免重复的攻击模式在“择优”过程的结果集中反复出现。实验证明,该方法能够高效准确地挖掘出网络中存在的复杂的攻击模式,实时对网络安全进行监控。

关键词: 攻击模式挖掘模型;最小二乘支持向量机;改进正交化算法;相关度检测法;改进滑动窗

中图分类号: TP393

文献标识码: A

文章编号: 1000-7180(2011)05-0171-05

Mining Model of Attacking Mode Based on Log Information

LONG Men¹, XIA Jing-bo¹, ZHANG Zi-yang²

(1 Telecommunication Engineering Institute, Air Force Engineering University, Xi'an 710077, China;

2 Engineering Institute, Air Force Engineering University, Xi'an 710038, China)

Abstract: This paper built the mining model of attacking mode to analyze log information combined the process “optimized selection” and “elimination”. The process “optimized selection” used the sparseness tactic based on Modified Classical Gram Schmidt (MCGS) to reset dataset, improved the efficiency of taxonomy model of Least Squares Support Vector Machines (LS-SVM). The “elimination” process eliminated the unreal attacking events, and presented the tactic of modified sliding-window based on correlated detection to avoid attacking modes repeating in the result set got from the process of “optimized selection”. The experiment validated this method can find out complex attacking modes in network efficiently and correctly.

Key words: mining model of attacking mode; LS-SVM; MCGS; correlated detection; modified sliding-window

1 引言

计算机网络中存在大量的主机、服务器和各种网络设备,在网络系统运行过程中,会产生大量的日志信息,包括系统日志、安全日志、应用日志和告警日志。这些日志之间存在一定的关联性,包含了安全事件的相关信息。通过对日志信息进行分析,可以挖掘出日志中包含的攻击事件。

文献[1]基于规则库方法对日志信息进行分析,主要通过规则匹配,准确率较高,但很难分析出未知安全事件;文献[2]设计并实现了一个基于关联规则的增量式模式挖掘算法,找出网络中频繁发生的行

为模式,自动建立异常和误用行为的模式库。但是该方法仅对攻击引起的消息进行分析,并没有结合网络的脆弱性。文献[3]把攻击行为所产生大量重复和相关报警中分类成不同超报警序列集合,从中挖掘频繁闭序列作为入侵场景,有效发现攻击过程。但该方法所生成的场景还仅仅是报警序列,导致场景集合中存在很多冗余信息。

文中提出了“择优”和“劣汰”过程结合的攻击模式挖掘模型。该模型分成两个阶段,“择优”过程是运用稀疏化最小二乘支持向量机对各种日志进行相关性分析,从复杂的日志信息中获取特定攻击事件发生的概率;第二阶段是“劣汰”过程,将攻击事件与相

收稿日期: 2010-07-20; 修回日期: 2010-08-22

基金项目: 陕西省自然科学基金(2009JM8001-1);军队武器装备科研项目(KJ * * *)

对应的脆弱性库进行比较,结合网络性能信息,剔除虚假的攻击事件,运用改进滑动窗策略修正得到节点安全态势和网络安全态势。

2 基于最小二乘支持向量机的“择优”过程

2.1 “择优”过程

Suykens J. A. K. 在 1999 年提出最小二乘支持向量机 (Least Squares Support Vector Machines, LS-SVM), 用于解决分类问题^[4]。该方法将原先的不等式约束转化为等式约束, 用求解线性规划方法代替支持向量机的二次规划方法, 提高了计算效率。对于一个个数为 l 的样本序列 $(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l), \dots, (x_l, y_l)$, 其输入向量为 $x_i \in R^n$, 输出向量为 $y_i \in R$, 采用最小二乘支持向量机进行函数估计, 优化问题可以描述如下:

$$\min \phi(w, b, e) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^l e_i^2 \quad (1)$$

$$\text{s. t. } \begin{cases} y_i = w^T \varphi(x_i) + b + e_i \\ i = 1, 2, \dots, l \end{cases} \quad (2)$$

其中, $\varphi(\cdot): R^n \rightarrow R^h$ 是将输入向量映射到高维特征空间的映射函数; 权值向量 $w \in R^h$, 误差变量 $e_i \in R$, 偏置值 $b \in R$; 惩罚系数 $\gamma > 0$, 从最小化目标函数可知, 惩罚系数的选择影响着误差所起的作用, 惩罚系数取值较大, 则误差影响占目标函数比重较大, 反之较小。

基于最小二乘支持向量机的分类模型的建立最终归根于求解模型系数, 即求解一个高维线性方程组 $H\theta = Y$, 其中实矩阵 $H = [h_1, h_2, \dots, h_n]$ 为 $n \times n$ 实矩阵, $n = l + 1$, 输出向量为 $Y = [0, y_1, y_2, \dots, y_l]^T$ 。最小二乘求解问题即可描述为求 $\theta = [\theta_1, \theta_2, \dots, \theta_n]^T$, 使得 $\|H\theta - Y\|$ 为最小, $\theta = [b, \alpha_1, \alpha_2, \dots, \alpha_l]^T$, α_i 对应为每一个训练样本的支持权重。

2.2 基于改进正交化算法的稀疏化算法

网络运行中产生的大量日志信息存在冗余信息, 这些冗余信息会影响日志信息的相关性分析结果的准确性。因此, 简化模型结构、剔除冗余信息是“择优”过程中亟待解决的问题。文中基于改进正交化算法 (Modified Classical Gram Schmidt, MCGS)^[5-6] 提出一种最小二乘支持向量机的新型稀疏化策略, 从 l 个训练样本中选择出 nsu 个最优样本, 约简支持向量。该方法将支持样本选择问题转化为高维映射核矩阵中输入向量对输出向量的贡献问题, 不需要迭代求解支持权重, 直接确认核心支持样本集, 实现稀疏化。

MCGS 算法基于选主元策略决定待正交化矩阵 H 的列向量的引入次序, 并同时以 CGS 算法对其实施正交化。记列重排后的矩阵为 H_c , 相应的正交化矩阵为 $Q_c = [q_1, q_2, \dots, q_n]$, 最小二乘解为 θ_c , 则 MCGS 算法如下描述:

Step1: 对于 $i = 1, 2, \dots, n$, 记 $q_1^{(i)} = h_i$, 计算

$$g_1^{(i)} = \langle q_1^{(i)}, Y \rangle / d_1^{(i)}; \quad (3)$$

$$[err]_1^{(i)} = (g_1^{(i)})^2 d_1^{(i)} / \langle Y, Y \rangle \quad (4)$$

在 $[err]_1^{(i)}, i = 1, 2, \dots, n$ 中选择最大值, 假定 $[err]_1^{(i_1)} = \max\{[err]_1^{(i)}, 1 \leq i \leq n\}$, 则 $q_1 = q_1^{(i_1)}$ 项被选作矩阵 Q_c 的第一列, 并以 $g_1 = g_1^{(i_1)}$ 作为 G_c 的第一个元素, 记 $[err]_1 = [err]_1^{(i_1)}$ 。记 $d_1 = d_1^{(i_1)} = \langle q_1, q_1 \rangle$, 相应的, 以 h_{i_1} 作为矩阵 H_c 的第一列, 仍记为 h_1 ; θ_{i_1} 作为相应最小二乘解 θ_c 第一个元素, 仍记为 θ_1 。记 $\text{Index}(1) = I_1$ 。

计算 $Y^{(1)} = Y - g_1 q_1$ 。

Step $k (k > 1)$: 对于 $i = 1, 2, \dots, n$, 且 $i \notin \{\text{index}(1), \text{index}(2), \dots, \text{index}(k-1)\}, j = 1, 2, \dots, k-1$, 执行如下算法:

$$r_{jk}^{(i)} = \langle q_j, h_i \rangle / d_j; \quad (5)$$

$$q_k^{(i)} = h_i - \sum_{j=1}^{k-1} r_{jk}^{(i)} q_j \quad (5)$$

$$d_k^{(i)} = \langle q_k^{(i)}, q_k^{(i)} \rangle;$$

$$g_k^{(i)} = \langle q_k^{(i)}, Y^{(k-1)} \rangle / d_k^{(i)} \quad (6)$$

$$[err]_k^{(i)} = (g_k^{(i)})^2 d_k^{(i)} / \langle Y, Y \rangle \quad (7)$$

同样对 $[err]_k^{(i)}, i \notin \{\text{index}(1), \text{index}(2), \dots, \text{index}(k-1)\}$ 进行最大值的筛选, 假设 $[err]_k^{(i_k)} = \max\{[err]_k^{(i)}, 1 \leq i \leq n, i \notin \{\text{index}(1), \text{index}(2), \dots, \text{index}(k-1)\}\}$, 则 $q_k = q_k^{(i_k)}$ (相应于矩阵 H 的 h_{i_k}), 被选作矩阵 Q_c 的第 k 列, 并以 $g_k = g_k^{(i_k)}$ 作为 G_c 的第 k 个元素, $[err]_k = [err]_k^{(i_k)}$ 。相应的, 以 h_{i_k} 作为矩阵 H_c 的第 k 列, 仍记为 H_{i_k} , 并以 θ_{i_k} 作为相应的最小二乘解 θ_c 的第 k 个元素, 同样仍记为 θ_k 。记 $d_k = d_k^{(i_k)} = \langle q_k, q_k \rangle$, 记 $\text{index}(k) = I_k$ 。计算 $Y^{(k)} = Y^{(k-1)} - g_k q_k$ 。

迭代过程在第 nsu 步终止, 满足 $J_N(nsuv) = \sqrt{\langle \epsilon, \epsilon \rangle / N} < \rho_{sl}$, 其中残差序列 $\epsilon = Y^{nsu}$, ρ_{sl} 为设定的理想阈值, 其值可以通过参考系统的信噪比及迭代过程中基于残差的损失函数稳定值来设定, N 为数据采样长度。当然, 在计算过程中如果某一迭代步骤上 $\langle q_k^{(i)}, q_k^{(i)} \rangle$ 小于某一预定较小值, 则相应项 h_i 将不被作为候选项, 以避免计算机编程中的

浮点溢出。

上面根据改进的 Gram-Schmidt 正交化算法将对特定攻击事件发生的概率影响较大的主要贡献项选择出来,针对最小二乘支持向量机而言,即从训练样本全集中选择主要支持向量,其编号包含于 $\{\text{index}(1), \text{index}(2), \dots, \text{index}(nsv)\}$ 。因为 $nsv \ll l$ 或 $nsv - 1 \ll l$ (偏置 b 对应项也选择出来),所以实现了稀疏化。使用优选样本再次训练,即可得到相对于原始样本的稀疏型最小二乘支持向量机。该策略优点在于全面考虑了原始样本对输出向量的影响,避免了剪枝算法中的多次训练过程。

3 基于滑动窗策略的“劣汰”过程

“择优”过程是通过分析网络的日志信息,得到特定攻击事件发生的概率。攻击事件是依照发生概率进行排序的,“择优”过程会选择前 N 个攻击事件作为结果,概率低的攻击事件会被排除在“择优”过程得到的结果之外。但是,攻击能否成功还取决于主机节点内部的脆弱性信息和网络性能信息。概率高的攻击事件有可能因为网络防范措施得当而成功,成为虚假攻击事件,而概率低的攻击事件则可能是因为包含了未知的攻击事件。

“劣汰”过程涉及到虚假攻击的剔除和新攻击事件的增补。对于得到的每一条攻击事件,首先检查是否包含了攻击所依赖的所有必要的漏洞,如果包含所有漏洞,则保持“择优”过程得到的结果;否则,将该攻击事件标志为虚假日志信息组合,进行剔除;将概率排名在 $N+1$ 的攻击事件吸纳进来,构成新的攻击事件集合,修正“择优”过程的模型,直到没有新的攻击事件补充。

文中提出一种基于相关度检测法的改进滑动窗策略,即基于优选样本的改进滑动窗策略。在保证训练样本长度不变的前提下,区别在于吸收过程中并非直接将最新样本吸纳,而是先判断该样本是否带来新的攻击相关联的信息。如果没有带来新信息,则保持评估模型不变;否则,将最新样本与已有的训练样本组成新训练样本集,并从中将虚假攻击事件的样本删除,保持训练样本容量的一致性。这样得到的模型更能反映网络安全评估的动态过程。

如何判定新样本是否带来新信息很重要,文中采取新样本与已有训练样本集的相关度进行比较的方法进行判断。如果新样本在映射空间能够被已有训练样本集描述,则新样本无需参训。样本在映射空间的相关描述定义如下:

新样本的输入仍为 x_{new} ,上一时刻的训练样本分别为 $x_i, i = 1, 2, \dots, nsv$;如果存在如下的描述关系,

$K(x, x_{\text{new}}) = \sum_{i=1}^{nsv} c_i K(x, x_i)$, c_i 为常系数,则新样本映射到高维映射空间后,线性依赖于 nsv 个 $(x_1, x_2, \dots, x_{nsv})$ 样本的支持向量,因此新样本对系统动态过程模型没有新影响,即其影响完全可以由上一时刻的训练样本描述。

在理论上采用上式计算新样本与已有训练样本集的线性相关性,有可能导致数值计算不稳定。因此,采用如下的近似条件判断:

$$\Delta_{\text{new}} = \|\varphi(x_{\text{new}}) - \sum_{i=1}^{nsv} c_i \varphi(x_i)\|^2 \leq \epsilon_{\text{new}} \quad (8)$$

其中, ϵ_{new} 为当前时刻界限值。因此,上式展开可描述如下:

$$\begin{aligned} \Delta_{\text{new}} &= \|\varphi(x_{\text{new}}) - \sum_{i=1}^{nsv} c_i \varphi(x_i)\|^2 \\ &= \varphi(x_{\text{new}}) \varphi(x_{\text{new}}) - 2 \varphi(x_{\text{new}}) \cdot \\ &\quad \left(\sum_{i=1}^{nsv} c_i \varphi(x_i) \right) + \sum_{i=1}^{nsv} c_i \sum_{j=1}^{nsv} c_j \varphi(x_j) \varphi(x_i) \\ &= K(x_{\text{new}}, x_{\text{new}}) - 2 \sum_{i=1}^{nsv} c_i K(x_i, x_{\text{new}}) + \\ &\quad \sum_{i=1}^{nsv} c_i \sum_{j=1}^{nsv} c_j K(x_i, x_j) \\ &\leq \epsilon_{\text{new}} \end{aligned} \quad (9)$$

最小化 Δ_{new} , 即可以求出训练样本对应的系数 $c_i, i = 1, 2, \dots, nsv$, 为了适当放宽影响,因此适当降低 c_i 的精度,在上式中加入 l_2 范数正则项 $\rho \|c\|^2$, 这样式(9)的最小化的左边写成:

$$\min K(x_{\text{new}}, x_{\text{new}}) - 2 \sum_{i=1}^{nsv} c_i K(x_i, x_{\text{new}}) + \sum_{i=1}^{nsv} c_i \sum_{j=1}^{nsv} c_j K(x_i, x_j) + \rho \|c\|^2$$

其中 ρ 为正的实常数。令 $[K]_{i,j} = K(x_i, x_j)$, $K_{\text{new}} = K(x_{\text{new}}, x_{\text{new}})$, $(K_{\text{new}})_i = K(x_{\text{new}}, x_i), i = 1, 2, \dots, nsv$, 则式(9)可以重写为

$$\min K_{\text{new}, \text{new}} - 2c^T \tilde{K}_{\text{new}} + c^T c \tilde{K} + \rho \|c\|^2 = c^T (\tilde{K} + \rho I) c - 2c^T \tilde{K}_{\text{new}} + K_{\text{new}, \text{new}} \quad (10)$$

$$c = (\tilde{K} + \rho I)^{-1} \tilde{K}_{\text{new}} \quad (11)$$

将式(11)代入式(10)可知:

$$\begin{aligned} \Delta &= K_{\text{new}, \text{new}} - \tilde{K}_{\text{new}}^T c \\ &= K_{\text{new}, \text{new}} - \tilde{K}_{\text{new}}^T ((\tilde{K} + \rho I)^{-1} \tilde{K}_{\text{new}}) \end{aligned} \quad (12)$$

因此,在每一步中计算出最小值,然后判断是否将样本加入到训练样本。将新样本与上一时刻训练样本集组合重新训练。这一“劣汰”过程既保证了得

到的攻击事件的准确性,又避免了遗漏未知的攻击事件.

4 实例分析

4.1 稀疏化“择优”算法的验证

为验证稀疏化“择优”算法的有效性,采用是 KDD99 数据集^[7]. 由于数据集过于庞大,因此文中依据各种类型数据在原始数据集中所占比例进行采样,这样可以保证实验数据集和原始数据集包含相等比例的各种攻击和正常网络数据. 训练集中包含的攻击如表 1 所示.

对上述数据集分别进行训练,然后依据最小二乘支持向量机“择优”算法进行测试输入的输出误差

对比. 取几个典型样本数,以稀疏化样本进行训练的模型称为 Optimize Model,而随机抽样训练所得到的模型称为 Rand Model,误差对比结果如表 2 所示.

表 1 训练集中包含的攻击

攻击类型	DoS 攻击	Probe 攻击	U2R 攻击	R2L 攻击
	smurf	port sweep	perl	warezmaster
攻	Neptune	satan	rootkit	imap
击	back	ip sweep	buffer_overflow	spy
形	teardrop	nmap		warezclient
式	land			ftp_write
	pod			phf

表 2 训练误差对比

样本数(/个)	29	39	49	59	69	79	89
Rand Model	1. 4756	1. 3695	0. 2415	0. 0292	0. 0191	0. 0133	0. 0122
Optimize Model	0. 5005	0. 4293	0. 0258	0. 0231	0. 0160	0. 0114	0. 0095

其中最典型的对比是在样本个数等于 49 时,稀疏化策略选择的样本训练的模型较之随机选择的样本训练产生的模型其误差有鲜明的对比,当然这种现象随着样本个数的递减而更加明显,随着样本个数的增加而趋于模糊.

4.2 “择优”和“劣汰”过程结合的模型验证

选取林肯实验室提供的 2000 年 DARPA 评估数据集作为实验数据,数据集中包含了网络边界和网络内部的数据信息^[8]. LLDOS1. 0 是 Lincoln 实验室在这两次攻击测试的基础上设计的一套多步攻击测试方案. 实验网络被划分为外网、DMZ 和内网三个区域,处于外网的攻击者通过对内网的三台主机进行入侵并操控进行 DDoS 攻击. 实验网络拓扑如图 1 所示.

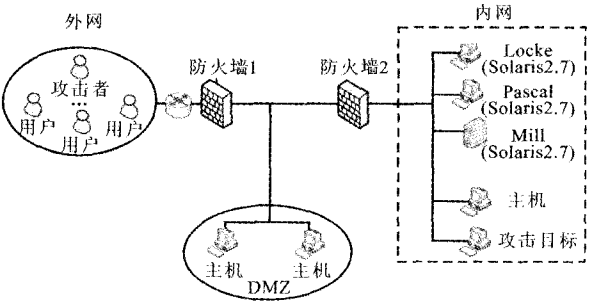


图 1 实验网络拓扑

运用稀疏化“择优”算法对攻击过程产生的日志信息进行分析,得到归一化的攻击概率,依据攻击概率列出排名前五的攻击,如表 3 所示.

表 3 “择优”结果集及其概率

顺序	攻击类型	攻击概率
1	Sadmind_Ping	0. 39
2	Stream_DoS	0. 31
3	Email_Almail_Overflow	0. 15
4	Rsh	0. 1
5	Sadmind_Amslverify_Overflow	0. 05

对“择优”算法中得到的攻击进行“劣汰”过程分析. Sadmind_Ping 的概率最高,运用“劣汰”算法,结合返回响应数据包的主机,确定出该次攻击已经成功,并且将返回响应数据包的主机确定为受到攻击的主机;对受到攻击的主机实施重点监控,将这些主机所运行的服务进行记录.

Email_Almail_Overflow 并不是一个单步攻击,分析发现仅与 Rsh 具有较高的概率函数依赖关系,而网络信息显示不包含使该攻击成功的漏洞,因此可以判定该攻击对网络安全不构成威胁. 网络设置中就已经说明此攻击为入侵者采用发送特征串的方法,引诱入侵检测系统发布大量的误报,掩盖真实的攻击行为. 因此,证明“择优”和“劣汰”过程结合的模型可以很好的剔除虚假攻击事件.

Sadmind_Amslverify_Overflow 的概率为 0. 05,而重点监控的主机中只有 Pascal、Locke 和 Mill 运行 sadmind 服务,运用“劣汰”算法,则可以判断

DDos 攻击极有可能在这三台主机上成功. Stream_DoS 的概率为 0.31.

该“择优”算法在分析日志信息时,未给出 Mstream_Zombie 的概率. 根据改进滑动窗策略的相关度检测法, Mstream_Zombie 与其它已知攻击的相关度小于设定的阈值 0.4, 因此, 将该攻击包含进新的训练样本集中, 重新进行训练, 得到的攻击概率排名如表 4 所示.

表 4 “择优”结果集及其概率

顺序	攻击类型	攻击概率
1	Sadmind_Ping	0.33
2	Stream_DoS	0.31
3	Sadmind_Amslverify_Overflow	0.2
4	Mstream_Zombie	0.15
5	Rsh	0.01

5 结束语

文中通过实例分析证明“择优”和“劣汰”过程结合的方法, 能够准确地挖掘网络中的攻击模式. 稀疏化的“择优”过程避免了冗余信息带来的计算误差, 提高了攻击模式挖掘的准确率. 而基于相关度的滑动窗“劣汰”过程有效地剔除了虚假攻击, 同时避免相同攻击模式反复出现, 优化了“择优”过程得到的结果集.

“择优”和“劣汰”过程结合的方法为网络安全评估模型的建立奠定了良好的基础, 将结果集与网络拓扑结构和节点重要性因子相结合, 就可以对网络安全进行准确评估, 此方向也是进一步的工作.

参考文献:

- [1] 韦勇, 连一峰. 基于日志审计与性能修正算法的网络安全态势评估模型[J]. 计算视学报, 2009, 32(4): 763-772.
- [2] 王丽苹, 安娜, 吴晓南, 等. 入侵检测系统中的行为模式挖掘[J]. 通信学报, 2004, 25(7): 168-175.
- [3] 郭帆, 涂风涛, 余敏. 基于报警序列的入侵场景自动构建[J]. 计算机应用, 2009, 29(8): 2223-2226.
- [4] Suykens J A K, Lukas L, Vandewalle J. Sparse approximation using least squares support vector machines [C]// Proceedings of IEEE Int symposium on circuits and system. Geneva, Switzerland: IEEE, 2000: 757-760.
- [5] Jiao L C, Bo L F, Wang L. Fast sparse approximation for least squares support vector machine [J]. IEEE Trans on Neural Networks, 2007, 18(3): 685-697.
- [6] Chen S, Chng E S, Alkadhimi K. Regularized orthogonal least squares algorithm for constructing radial basis function networks[J]. International Journal of Control, 1996, 64(5): 821-837.
- [7] KDD-99. KDD cup 1999 data [EB/OL]. (2006). [2010-03-13]. <http://kdd.ics.uci.edu/databases/kdd-cup99/kddcup99.html>.
- [8] MIT Lincoln Lab. 2000 DARPA intrusion detection scenario specific datasets [EB/OL]. (2003) [2010-03-13] http://www.ll.mit.edu/IST/ideval/data/2000/2000_data_index.html, 2003-7.

作者简介:

龙 门 女, (1982-), 博士研究生. 研究方向为网络信息安全与管理.

(上接第 170 页)

- [4] Chang L C, Ton L R, Kao M F, et al. Stack operations folding in Java processors[D]. Taiwan: National Chiao Tung University, 1998.
- [5] 于海, 樊晓桢. 基于 FPGA 异步 FIFO 的研究与实现[J]. 微电子学与计算机, 2007, 24(3): 210-216.
- [6] 赵刚, 侯立刚, 刘源, 等. 基于 SoC 设计的软硬件协同验证方法学[J]. 微电子学与计算机, 2006, 22(6): 24-26.

作者简介:

胡 平 男, (1962-), 副教授, 硕士生导师. 研究方向为嵌入式系统、计算智能.

张金钟 男, (1984-), 硕士. 研究方向为嵌入式系统、信号处理.