一种基于压缩边界Fisher分析的硬件木马检测方法

王晓晗 王 韬 李雄伟* 张 阳 黄长阳
(陆军工程大学石家庄校区装备模拟训练中心 石家庄 050003)

摘 要：针对物理环境下旁路分析技术对电路中规模较小的硬件木马检出率低的问题，该文引入边界Fisher分析(MFA)方法，并提出一种基于压缩边界Fisher分析(CMFA)的硬件木马检测方法。通过减小样本的同类近邻样本与该样本以及类中心之间距离和增大类中心的同类近邻样本与异类样本之间距离的方式，构建投影空间，发现原始功耗旁路信号中的差异特征，实现硬件木马检测。AES加密电路中的硬件木马检测实验表明，该方法具有比已有检测方法更高的检测精度，能够检测出占原始电路规模0.04%的硬件木马。

关键词：硬件木马检测；集成电路；旁路分析；流形学习；边界Fisher分析
中图分类号：TN918 **文献标识码**：A **文章编号**：1009-5896(2019)12-3043-08
DOI: 10.11999/JEIT190004

A Hardware Trojan Detection Method Based on Compression Marginal Fisher Analysis

WANG Xiaohan WANG Tao LI Xiongwei ZHANG Yang HUANG Changyang
(*Equipment Simulation Training Center*, *Army Engineering University Shijiazhuang Campus*, *Shijiazhuang* 050003, *China*)

Abstract: Against the problem of low detection rate to detect small hardware Trojan by side-channel in physical environment, the Marginal Fisher Analysis (MFA) is introduced. On the basis, a hardware Trojan detection method based on Compression Marginal Fisher Analysis (CMFA) is proposed. The projection space is constructed by reducing the distance between the sample and its same neighbor samples, and the distance between the same neighbor samples and the center of the same kind, and increasing the distance between the same neighbor samples of the center and the sample in different kind. Thus, the difference in the original data is found without any assumptions about data distribution, and the detection of hardware Trojan is achieved. The hardware Trojan detection experiment in AES encryption circuit shows that this method can effectively distinguish the statistical difference in side-channel signal between reference chip and Trojan chip and detect the hardware Trojan whose scale is 0.04% of the original circuit.

Key words: Hardware Trojan detection; Integrated Circuit(IC); Side-channel analysis; Manifold learning; Marginal Fisher Analysis (MFA)

1 引言

随着微电子学与计算机技术的快速发展，能够保存和处理敏感信息的集成电路(Integrated Circuit, IC)已被广泛地应用于人们日常生活的各个领域，影响着人们的生活方式。然而，由于当前IC芯片的设计与制造分离，IC芯片的安全难以保证，很容易被植入具有恶意功能的冗余电路——硬件木马[1](Hardware Trojan, HT)。硬件木马能够在特定情况下破坏IC芯片和泄露秘密信息，一旦被不法分子利用，将严重威胁个人利益甚至是国家安全，因此加强IC芯片中的硬件木马检测，保证IC芯片安全具有重要意义。

目前，对芯片内硬件木马进行检测的主流方法可分为逻辑测试和旁路分析两类[2]。其中逻辑测试主要是生成测试向量，并比对“金片”(确定无硬件木马的芯片，可先对芯片进行足够多的I/O测试并记录其工作时的相关信息作为参考，再通过剖片确认芯片设计中无硬件木马，下同)与待测IC的输出结果是否一致来判断芯片中是否含有硬件木马。例如，文献[3]以多次触发惰性节点(电路中0-1翻转

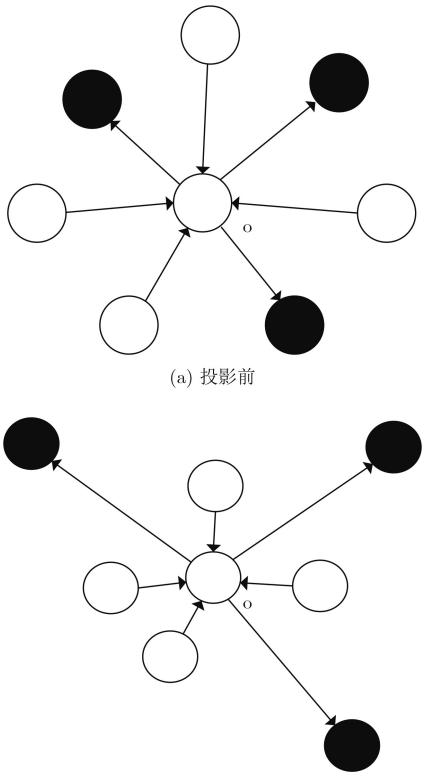
收稿日期：2019-01-03；改回日期：2019-03-14；网络出版：2019-05-28
*通信作者： 李雄伟 lxw-wys@163.com
基金项目：国家自然科学基金(61602505)
Foundation Item: The National Natural Science Foundation of China (61602505)

概率低的节点)的方式生成测试向量集。在此基础上，文献[4]利用遗传算法生成测试向量以激活电路中更难被触发的硬件木马。文献[5]针对多输入电路生成完备的测试向量集以激活电路输入中潜藏的硬件木马。文献[6]对电路进行分区并针对电路结构生成最优的测试向量，以增大硬件木马对电路功耗的影响。然而这类方法对大规模电路的分析代价较高，需要分析人员熟悉电路内部细节。旁路分析主要是比对相同条件下测得的“金片”和待测IC的旁路特征参数(功耗[7]、电磁[8]、延迟[9]等)来检测芯片中的硬件木马。例如，文献[10]以IC仿真的旁路特征参数作为训练样本，检测生产后芯片中的硬件木马。文献[11]通过测量环境改变后DC直流传输特性参数的变化以检测电路中的热载流子注入型硬件木马。为克服噪声的影响，一些研究采用特征变换方法处理旁路信号以达到检测的目的。例如文献[12]在信息损失尽量小的前提下将旁路信号投影到低维子空间以检测硬件木马，能够检测出占电路规模10%的硬件木马。文献[13]在旁路信号统计模型的基础上，先将旁路信号投影到高维空间增加可分性，再投影到低维子空间，以减少噪声对检测的影响，能够检测出占电路规模2%的硬件木马。文献[14]采用自组织竞争神经网络在尽量不损失有用信息的前提下检测出占电路规模0.16%的硬件木马。相比于逻辑测试而言，旁路分析更简单、便捷，但检测效果易受噪声干扰，依赖特征变换方法的选取。

基于功耗旁路分析的硬件木马检测本质是旁路信号的分类问题。如果硬件木马的规模较大，则会显著改变旁路信号的旁路特征，而设计精良的硬件木马对宿主电路的改变很小，其对旁路信号的影响亦很小。检测时，噪声干扰是影响检测效果的主要因素：一方面噪声掩盖了硬件木马对旁路信号的影响，导致检出率下降；另一方面噪声是由环境噪声、实验设备噪声、集成电路的工艺噪声以及集成电路运算时产生的转换噪声等多种噪声耦合而成，导致物理环境下采集到的旁路信号分布十分复杂。此时，采用高斯分布等理想模型对旁路信号集进行刻画可能会损失部分有效信息，利用K-L变换[7]和核MMC[13]等依赖高斯分布建模的方法对小规模硬件木马的检测效果并不理想。因此本文引入流形学习方法边界Fisher分析(Marginal Fisher Analysis, MFA)进行硬件木马检测，并在其几何原理的基础上，提出了一种适用于硬件木马检测的压缩边界Fisher分析(Compression Marginal Fisher Analysis, CMFA)方法。该方法可在不对数据进行任何假设的情况下，发现数据间的规律，有效提取“金片”和含硬件木马IC旁路信号的差异信息，实现硬件木马检测。

2 边界Fisher分析(MFA)

MFA是基于图嵌入框架的有监督流形学习方法[15]，在降维过程中不需要假设样本的分布，其基本思想是：高维空间中离得近的同类样本映射到低维空间中也应保持接近，而离得近的异类样本映射到低维空间中应当离得远。如图1所示，颜色相同的节点表示同类近邻样本，颜色不相同的节点表示异类近邻样本，则投影后样本点o与其异类近邻样本的距离增大，与同类近邻样本的距离减小。



(a) 投影前

(b) 投影后

图 1 MFA算法的基本思想

对于给定的样本集$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}\in R^{D\times N}$，MFA算法寻找最优投影矩阵$\boldsymbol{\Phi}=\{\varphi_1,\varphi_2,\cdots,\varphi_d\}$以及样本在低维空间的投影$\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_N\}\in R^{d\times N}$(空间维数$d<<D$)，使得$\boldsymbol{y}_i=\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{x}_i$，具体步骤为：

(1)构建邻接图G：对于样本集$\boldsymbol{X}$中的每个样本$\boldsymbol{x}_i$，算法定义本征图G^{B}和惩罚图G^{P}分别描述该样本的k^{B}个同类近邻样本和k^{P}个异类近邻样本(一般利用K-近邻法或ε邻域法判断是否为近邻样本)，图的顶点代表样本点，若样本$\boldsymbol{x}_j$为$\boldsymbol{x}_i$的近邻样本，则用边将其连接；

(2)计算权重系数矩阵$\boldsymbol{W}$：根据本征图G^{B}和惩罚图G^{P}，分别利用$\boldsymbol{W}_{ij}{}^{\mathrm{B}}$和$\boldsymbol{W}_{ij}{}^{\mathrm{P}}\in\{0,1\}$表示图中边的权重，若样本$\boldsymbol{x}_i$和$\boldsymbol{x}_j$不相连，则$\boldsymbol{W}_{ij}$为0，否则$\boldsymbol{W}_{ij}$为1；

(3)计算投影矩阵$\boldsymbol{\Phi}$：由权重系数矩阵$\boldsymbol{W}$重构低维空间中的样本投影$\boldsymbol{y}_i$，其优化准则定义为

$$\begin{aligned}\max & \frac{\sum_{i,j=1}^{N}\left\|\boldsymbol{y}_i-\boldsymbol{y}_j\right\|^2\boldsymbol{W}_{ij}^{\mathrm{P}}}{\sum_{i,j=1}^{N}\left\|\boldsymbol{y}_i-\boldsymbol{y}_j\right\|^2\boldsymbol{W}_{ij}^{\mathrm{B}}}\\ &=\frac{\sum_{i,j=1}^{N}\left\|\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{x}_j\right\|^2\boldsymbol{W}_{ij}^{\mathrm{P}}}{\sum_{i,j=1}^{N}\left\|\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{x}_i-\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{x}_j\right\|^2\boldsymbol{W}_{ij}^{\mathrm{B}}}\\ &=\frac{\mathrm{tr}\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^{\mathrm{P}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Phi}\right)}{\mathrm{tr}\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}^{\mathrm{B}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Phi}\right)}\end{aligned}\tag{1}$$

其中，$\boldsymbol{L}^{\mathrm{B}}$和$\boldsymbol{L}^{\mathrm{P}}$分别为本征图$G^{\mathrm{B}}$和惩罚图$G^{\mathrm{P}}$的拉普拉斯矩阵，$\boldsymbol{L}^{\mathrm{B}}=\boldsymbol{D}^{\mathrm{B}}-\boldsymbol{W}^{\mathrm{B}}$, $\boldsymbol{L}^{\mathrm{P}}=\boldsymbol{D}^{\mathrm{P}}-\boldsymbol{W}^{\mathrm{P}}$, $\boldsymbol{D}^{\mathrm{B}}$和$\boldsymbol{D}^{\mathrm{P}}$均为对角矩阵，$\boldsymbol{D}_{ii}^{\mathrm{B}}=\sum_j\boldsymbol{W}_{ij}^{\mathrm{B}}$, $\boldsymbol{D}_{ii}^{\mathrm{P}}=\sum_j\boldsymbol{W}_{ij}^{\mathrm{P}}$。此时式(1)可转换为广义特征值求解问题[15]，即

$$\boldsymbol{X}\boldsymbol{L}^{\mathrm{B}}\boldsymbol{X}^{\mathrm{T}}\varphi_i=\lambda_i\boldsymbol{X}\boldsymbol{L}^{\mathrm{P}}\boldsymbol{X}^{\mathrm{T}}\varphi_i,\ i=1,2,\cdots,d\tag{2}$$

其中，φ_i为特征值最大的前d个广义特征向量(投影方向)，$\boldsymbol{y}_j(i)$则为样本$\boldsymbol{x}_j$在φ_i上的投影。

3　压缩边界Fisher分析

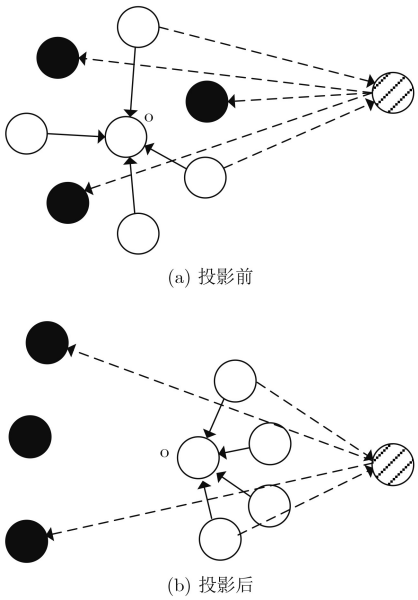
MFA算法虽然在构建投影空间时约束同类近邻样本离得近，异类近邻样本离得远，却不能保证投影后的样本有利于分类，如图1所示，黑色样本虽然远离白色样本o但并未与其它白色样本分离。因此本文提出了一种CMFA方法，以每类样本的类中心为基准点，通过本征图G^{B}、压缩图G^{C}和内核差异图G^{CD}构建投影空间，使同类近邻样本离得近的同时更靠近同类基准点，异类样本与基准点离的尽可能远。该算法思想如图2所示，阴影点为基准点，虚线为基准点对样本的影响，显然加入基准点后更有利于区分两类样本，提高硬件木马检测效果。

与MFA相同，本征图主要描述样本与其同类近邻样本与之间的邻接关系，其权重矩阵$\boldsymbol{W}^{\mathrm{B}}$为

$$\boldsymbol{W}_{ij}^{\mathrm{B}}=\begin{cases}1, & j\in N_{k^{\mathrm{B}}}(i)\vee i\in N_{k^{\mathrm{B}}}(j)\\ 0, & \text{其它}\end{cases}\tag{3}$$

其中，$N_{k^{\mathrm{B}}}(i)$为与样本点$\boldsymbol{x}_i$属于同一类的k^{B}个最近邻样本的下标集合。为了使投影空间中同类近邻样本离得近，其函数约束式为

$$\min\sum_{i,j=1}^{N}\left\|\boldsymbol{y}_i-\boldsymbol{y}_j\right\|^2\boldsymbol{W}_{ij}^{\mathrm{B}}\tag{4}$$



(a) 投影前

(b) 投影后

图 2 CMFA算法的基本思想

压缩图G^{C}主要描述样本与其同类近邻样本之间的邻接关系，其目的是使同类近邻样本在投影空间中离类中心更近，以增大同类样本的紧致性，为了使同类近邻样本在投影空间中更靠近类中心，将权重矩阵$\boldsymbol{W}^{\mathrm{C}}$定义为

$$\boldsymbol{W}_{ij}^{\mathrm{C}}=\begin{cases}\dfrac{(\boldsymbol{m}_k-\boldsymbol{x}_i)(\boldsymbol{x}_j-\boldsymbol{x}_i)}{\sqrt{\left\|\boldsymbol{m}_k-\boldsymbol{x}_i\right\|^2}\cdot\sqrt{\left\|\boldsymbol{x}_j-\boldsymbol{x}_i\right\|^2}}, & \\ \quad k\in l(i), j\in N_{k^{\mathrm{C}}}(i) & \\ \dfrac{(\boldsymbol{m}_k-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)}{\sqrt{\left\|\boldsymbol{m}_k-\boldsymbol{x}_j\right\|^2}\cdot\sqrt{\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|^2}}, & \\ \quad k\in l(j), i\in N_{k^{\mathrm{C}}}(j) & \\ 0, \text{其它} & \end{cases}\tag{5}$$

其中，$l(i)$为$\boldsymbol{x}_i$的类别标号，$\boldsymbol{m}_k$为$\boldsymbol{x}_i$所在类的样本均值，$N_{k^{\mathrm{C}}}(i)$为与样本点$\boldsymbol{x}_i$属于同一类的k^{C}个最近邻样本的下标集合。压缩图的直观解释如图3所示，b为样本的类中心，样本a为样本o的同类近邻样本，则$\boldsymbol{W}_{\mathrm{oa}}^{\mathrm{C}}$为直线oa和ob之间夹角$\theta$的余弦值。a′为样本a在直线ob上的投影，为了使样本a更靠近类中心b，只需调整a′与o之间的距离即可。当$\theta\leq90°$时，$\boldsymbol{W}_{\mathrm{oa}}^{\mathrm{C}}\geq0$, oa′ =oa$\cdot\boldsymbol{W}_{\mathrm{oa}}^{\mathrm{C}}$，增大oa$\cdot\boldsymbol{W}_{\mathrm{oa}}^{\mathrm{C}}$将使a′离b更近；当$\theta>90°$时，$\boldsymbol{W}_{\mathrm{oa}}^{\mathrm{C}}<0$, -oa′=oa$\cdot\boldsymbol{W}_{\mathrm{oa}}^{\mathrm{C}}$，增大oa$\cdot\boldsymbol{W}_{\mathrm{oa}}^{\mathrm{C}}$将使a′离b更近。因此，为了使近邻样本更靠近基准点，增大同类样本的紧致性，函数约束式定义为

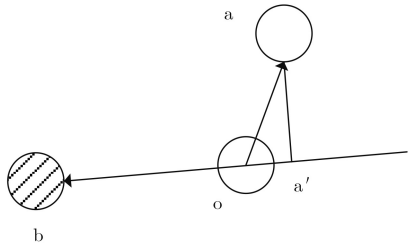


图 3 压缩图原理分析

$$\max \sum_{i,j=1}^{N} \left\| \boldsymbol{y}_i - \boldsymbol{y}_j \right\|^2 W_{ij}^{\mathrm{C}} \tag{6}$$

内核差异图G^{CD}主要是利用类均值样本的近邻样本集表征类中心，将与样本异类的类中心的同类近邻样本作为样本的异类邻接样本，通过增大样本与异类类中心的近邻样本集之间的距离，使投影空间中两类样本分离，具体做法是选定两个类中心的近邻样本集ξ_1和ξ_2，对于样本集ξ_1所在类的每个样本而言，其异类邻接样本为样本集ξ_2中的所有样本，反之亦然，则权重矩阵$\boldsymbol{W}^{\mathrm{CD}}$定义如下

$$\boldsymbol{W}_{ij}^{\mathrm{CD}} = \begin{cases} 1, & j \in N_{k^{\mathrm{CD}}}(i) \vee i \in N_{k^{\mathrm{CD}}}(j) \\ 0, & \text{其它} \end{cases} \tag{7}$$

其中，$N_{k^{\mathrm{CD}}}(i)$为与$\boldsymbol{x}_i$异类的类均值样本的k^{CD}个同类近邻样本的下标集合。内核差异图的直观解释如图4所示，点o为4个白色样本的类中心，4个白色样本为o的近邻样本，对于任一黑色样本而言，其异类邻接样本为点o的所有同类近邻样本，当样本远离其异类邻接样本时，两类样本间的距离增大，因此，其函数约束式定义为

$$\max \sum_{i,j=1}^{N} \left\| \boldsymbol{y}_i - \boldsymbol{y}_j \right\|^2 \boldsymbol{W}_{ij}^{\mathrm{CD}} \tag{8}$$

综合式(4)、式(6)和式(8)的优化目标，CMFA的优化准则定义为

$$\begin{aligned} &\max \frac{\sum_{i,j=1}^{N} \left\| \boldsymbol{y}_i - \boldsymbol{y}_j \right\|^2 \boldsymbol{W}_{ij}^{\mathrm{CD}} + \sum_{i,j=1}^{N} \left\| \boldsymbol{y}_i - \boldsymbol{y}_j \right\|^2 \boldsymbol{W}_{ij}^{\mathrm{C}}}{\sum_{i,j=1}^{N} \left\| \boldsymbol{y}_i - \boldsymbol{y}_j \right\|^2 W_{ij}^{\mathrm{B}}} \\ &= \frac{\mathrm{tr}\left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{X} (\boldsymbol{L}^{\mathrm{CD}} + \boldsymbol{L}^{\mathrm{C}}) \boldsymbol{X}^{\mathrm{T}} \boldsymbol{\Phi} \right)}{\mathrm{tr}\left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{L}^{\mathrm{B}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{\Phi} \right)} \end{aligned} \tag{9}$$

其中$\boldsymbol{L}^{\mathrm{CD}}$, $\boldsymbol{L}^{\mathrm{C}}$, $\boldsymbol{L}^{\mathrm{B}}$分别为图$G^{\mathrm{CD}}$, G^{C}和G^{B}的拉普拉斯矩阵(计算方式同上)，同样式(9)可转换为广义特征值求解问题，即

$$\boldsymbol{X} \boldsymbol{L}^{\mathrm{B}} \boldsymbol{X}^{\mathrm{T}} \varphi_i = \lambda_i \boldsymbol{X} (\boldsymbol{L}^{\mathrm{CD}} + \boldsymbol{L}^{\mathrm{C}}) \boldsymbol{X}^{\mathrm{T}} \varphi_i, \ i = 1, 2, \cdots, d \tag{10}$$

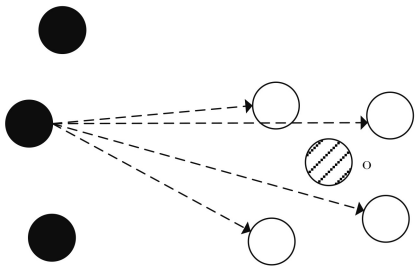


图 4 内核差异图原理分析

φ_i即为CMFA的前d个特征值最大的广义特征向量(投影方向)。

4 理论分析

对离散旁路信号进行变换的本质是利用特征提取方法寻找一组正交投影方向(特征向量)，并计算旁路信号在每个投影方向上的投影，只要某个方向上的投影存在差异则判定含有硬件木马，因此可通过对比最优投影方向φ上的样本投影情况判断检测效果。由于CMFA和MFA方法的函数约束式均为Rayleigh熵形式，两种算法下投影方向φ的判别能力分别为[16]

$$\begin{aligned} \rho_{\mathrm{MFA}}(\varphi) &= \frac{\sum_{i,j=1}^{N} \left\| \varphi^{\mathrm{T}} \boldsymbol{x}_i - \varphi^{\mathrm{T}} \boldsymbol{x}_j \right\|^2 \boldsymbol{W}_{ij}^{\mathrm{P}}}{\sum_{i,j=1}^{N} \left\| \varphi^{\mathrm{T}} \boldsymbol{x}_i - \varphi^{\mathrm{T}} \boldsymbol{x}_j \right\|^2 \boldsymbol{W}_{ij}^{\mathrm{B}}} \\ &= \frac{\varphi^{\mathrm{T}} \sum_{i,j=1}^{N} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \boldsymbol{W}_{ij}^{\mathrm{P}} \varphi}{\varphi^{\mathrm{T}} \sum_{i,j=1}^{N} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \boldsymbol{W}_{ij}^{\mathrm{B}} \varphi} \end{aligned} \tag{11}$$

$$\begin{aligned} &\rho_{\mathrm{CMFA}}(\varphi) \\ &= \frac{\varphi^{\mathrm{T}} \sum_{i,j=1}^{N} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \left(\boldsymbol{W}_{ij}^{\mathrm{CD}} + \boldsymbol{W}_{ij}^{\mathrm{C}} \right) \varphi}{\varphi^{\mathrm{T}} \sum_{i,j=1}^{N} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \boldsymbol{W}_{ij}^{\mathrm{B}} \varphi} \end{aligned} \tag{12}$$

由于式(11)与式(12)分母的表达式相同，此处仅比较分子即可。类似于Fisher判别分析的表达式$\frac{\varphi^{\mathrm{T}}(\boldsymbol{S}_b)\varphi}{\varphi^{\mathrm{T}}(\boldsymbol{S}_w)\varphi}$，其类间离散度矩阵$\boldsymbol{S}_b$的迹越大，表示投影后两类样本的类中心离得远，类内离散度矩阵$\boldsymbol{S}_w$的迹越小，表示投影后样本与同类类中心离得近。则分子的迹越大，分母的迹越小，ρ的值越大，所求投影方向φ的分类性能越好。同理，对于

给定的样本集$\boldsymbol{X}$，分子中的$\mathrm{tr}\left(\sum_{i,j=1}^{N}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{W}_{ij}\right)$越大，算法计算出来的投影方向$\varphi$的分类性能就越好，其中$(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}$表示样本间的离散度矩阵，其迹$\mathrm{tr}\left((\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\right)=\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2$。因为

$$\sum_{j=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\boldsymbol{W}_{ij}^{\mathrm{P}}=l_1^{\mathrm{P}}+l_2^{\mathrm{P}}+\cdots+l_{k^{\mathrm{P}}}^{\mathrm{P}} \tag{13}$$

$$\begin{aligned}&\sum_{j=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\left(\boldsymbol{W}_{ij}^{\mathrm{CD}}+\boldsymbol{W}_{ij}^{\mathrm{C}}\right)\\&\quad=\left(l_1^{\mathrm{C}}+l_2^{\mathrm{C}}+\cdots+l_{k^{\mathrm{C}}}^{\mathrm{C}}\right)\\&\qquad+\left(l_1^{\mathrm{CD}}+l_2^{\mathrm{CD}}+\cdots+l_{k^{\mathrm{CD}}}^{\mathrm{CD}}\right)\end{aligned} \tag{14}$$

其中，l_j^{P}，l_j^{CD}和l_j^{C}分别表示样本$\boldsymbol{x}_j$与样本$\boldsymbol{x}_i$之间的距离，根据$\boldsymbol{W}_{ij}$所描述的样本之间的邻接关系，l_j^{P}为样本$\boldsymbol{x}_i$的k^{P}个异类近邻样本与$\boldsymbol{x}_i$的距离和，l_j^{CD}为与样本$\boldsymbol{x}_i$异类的类中心的k^{CD}个同类近邻样本与$\boldsymbol{x}_i$的距离和，显然当$k^{\mathrm{CD}}=k^{\mathrm{P}}$时，$\sum_{j=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\boldsymbol{W}_{ij}^{\mathrm{P}}<\sum_{j=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\boldsymbol{W}_{ij}^{\mathrm{CD}}$。则当$k^{\mathrm{CD}}=k^{\mathrm{P}}$时，若存在$k^{\mathrm{C}}$使得$\sum_{i,j=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\boldsymbol{W}_{ij}^{\mathrm{C}}>\sum_{i,j=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\boldsymbol{W}_{ij}^{\mathrm{P}}-\sum_{i,j=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\boldsymbol{W}_{ij}^{\mathrm{CD}}$成立，CMFA方法的分类性能优于MFA。

5　硬件木马检测方案

结合硬件木马检测的一般流程[12]，下面给出基于CMFA的硬件木马检测方案步骤(基于MFA的硬件木马检测方案与之类似)：

步骤 1　旁路信号的采集。在同一测试向量下采集“金片”和待测IC的功耗旁路信号，得到旁路信号集$\boldsymbol{X}=\boldsymbol{L}_1\cup\boldsymbol{L}_2$，其中，$\boldsymbol{L}_1=\{b(i,j)|i=1,2,\cdots,N/2;\ j=1,2,\cdots,M\}$和$\boldsymbol{L}_2=\{t(i,j)|i=1,2,\cdots,N/2;\ j=1,2,\cdots,M\}$分别为“金片”和待测IC的功耗旁路信号集，$N$为旁路信号个数，$M$为每条旁路信号的采样长度。

步骤 2　计算权重矩阵。设定近邻参数k^{B}, k^{C}和k^{CD}并根据式(3)、式(5)、式(7)分别计算本征图、压缩图和内核差异图的权重矩阵$\boldsymbol{W}^{\mathrm{B}}$, $\boldsymbol{W}^{\mathrm{C}}$和$\boldsymbol{W}^{\mathrm{CD}}$。

步骤 3　计算拉普拉斯矩阵。根据步骤2中的权重矩阵分别计算相应的拉普拉斯矩阵$\boldsymbol{L}^{\mathrm{B}}$, $\boldsymbol{L}^{\mathrm{C}}$和$\boldsymbol{L}^{\mathrm{CD}}$。

步骤 4　计算特征向量集。根据式(10)以及步骤3中的拉普拉斯矩阵计算出特征向量集$\boldsymbol{\Phi}$以及旁路信号集$\boldsymbol{L}_1$和$\boldsymbol{L}_2$在投影方向上的投影$\boldsymbol{L}_1'$和$\boldsymbol{L}_2'$。

步骤 5　通过比对信号集$\boldsymbol{L}_1'$和$\boldsymbol{L}_2'$的差异来判断是否含有木马。若不能判断待测IC中含有硬件木马则重复步骤2直至检测出硬件木马或遍历所有的参数取值为止(此时认为待测芯片中没有硬件木马)。

6　实验验证

6.1　实验配置

采用配有XC5VLX30 FPGA芯片的SASEBO旁路攻击标准评估板进行检测实验，PC端使用LabVIEW编写的虚拟仪器控制平台控制整个采集过程，示波器Tektronix DPO4104用于采集FPGA工作时的功耗旁路信号，USB数据线用于数据传输。实验时在FPGA中分别运行含硬件木马和不含硬件木马的AES(Advanced Encryption Standard)加密电路，其逻辑规模为4453个6输入LUT和1270个寄存器。实验中植入的硬件木马为组合型，分别为32位比较器(木马1)和8位比较器(木马2)，在特定明文条件下才会被激活。两种木马的逻辑规模分别为5个LUT和2个LUT，分别占原始AES电路的0.11%和0.04%。对于同规模的其他类型硬件木马而言，本文方法同样适用。采集时使用固定明文进行采样，采样频率设为2.50 GSa/s，采样时长为0.4 μs，每1000个采样点作为一条旁路信号，分别采得“金片”和含木马电路的功耗旁路信号各1000条，作为后期处理的样本集，为了削弱噪声的影响，每条旁路信号均为16次采样求平均的结果。

6.2　实验结果分析

为了验证CMFA方法的有效性，以每个投影方向上“金片”样本投影的$\mu\pm3\sigma$(σ表示标准差，下同)为检测边界，以简单投票法[17]为判别标准，即统计待测IC在每个投影方向上落在$\mu\pm3\sigma$范围内的样本数与范围外的样本数，以落于边界外样本数最大的特征向量为最优投影方向，若该特征向量上位于边界范围外的样本数多于范围内的样本数则判定含有硬件木马。由于实验对1000条待测IC的旁路信号进行分析，因此检测阈值设定为500，超出检测边界的样本数多于500则判定含有硬件木马。实验中近邻数k的取值影响检测效果，但目前尚无有效理论确定近邻数k的最优取值，一般根据经验从2～20之间进行选取[16]，本文MFA方法的参数为$k^{\mathrm{B}}=k^{\mathrm{P}}=15$, CMFA的参数为$k^{\mathrm{B}}=k^{\mathrm{C}}=k^{\mathrm{CD}}=15$。

图5为采用K-L变换[7]，核MMC[13]，MFA和CMFA对木马1的检测结果，横坐标表示按照特征值大小排序的特征向量，纵坐标为每个特征向量上的投影值，深色部分为“金片”的投影分布，浅色部分为含木马IC的投影分布，两条黑色虚线为检测

边界。直观来看，MFA和CMFA在第1个特征向量上有良好的区分效果，大多数浅色样本超出检测边界，经统计，超出3σ范围的样本投影数分别为930和992，能检测出硬件木马，而核MMC(第1个特征向量)和K-L变换(第2个特征向量)仅有少部分浅色样本超出边界，经统计超出3σ范围的样本投影数分别为235个和147个，均不能判断是否含有木马。

图6为采用4种方法对木马2的检测结果。此

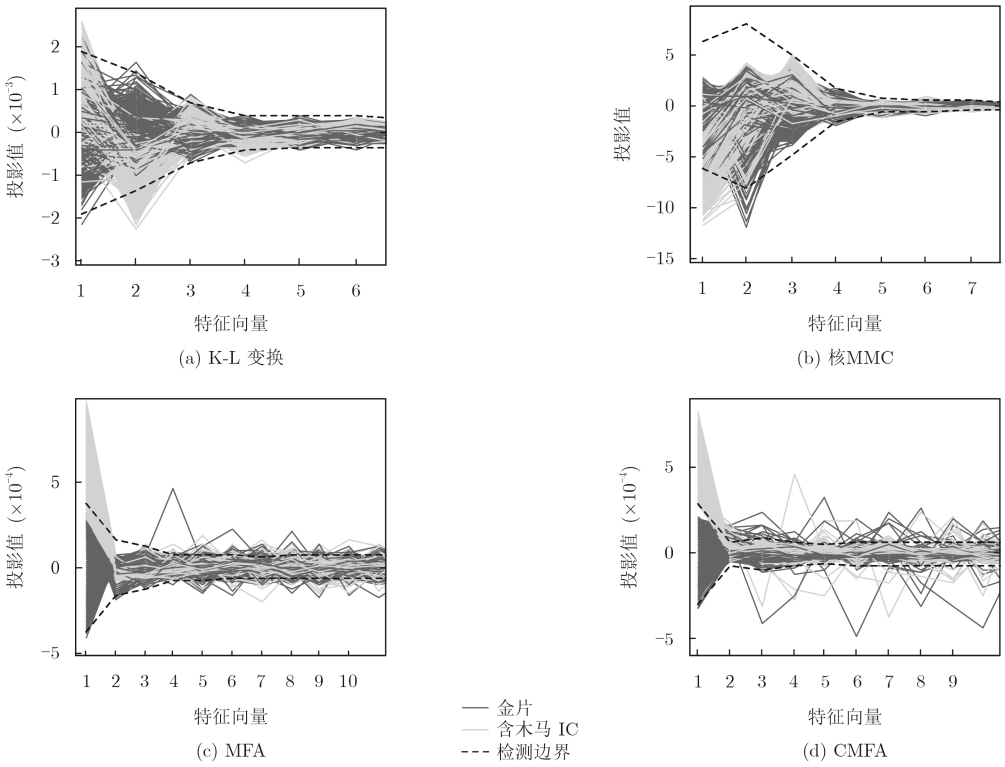


图 5 采用K-L变换, 核MMC, MFA和CMFA对木马1的检测结果

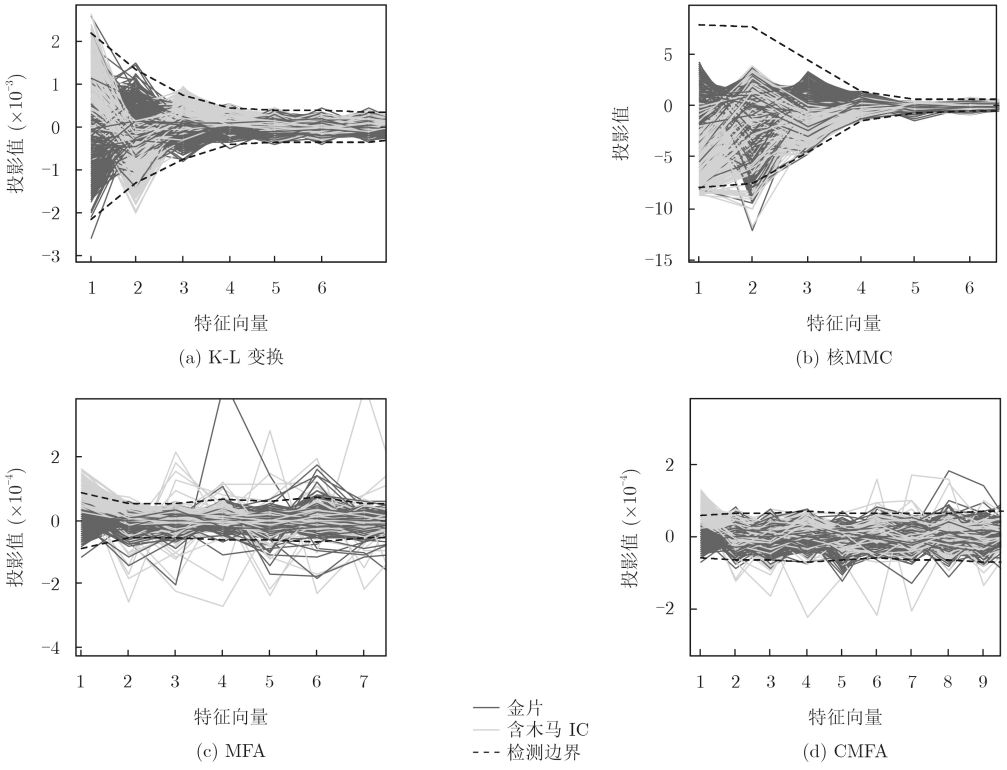


图 6 采用K-L变换, 核MMC, MFA和CMFA对木马2的检测结果

时，核MMC和K-L变换在最优投影方向上超出检测边界的样本数分别为43个(第1个特征向量)和36个(第2个特征向量)，均不能判断出硬件木马，而MFA和CMFA在第1个特征向量上测得检测边界外的样本数分别为429个和701个，仅CMFA能检测出硬件木马。

从图5和图6中可以看出，在第1个投影方向上CMFA方法的同类样本投影范围(两个检测边界之间的宽度)分别为5×10^{-4}和10^{-4}，均小于MFA方法的8×10^{-4}和2×10^{-4}, CMFA使投影空间中同类样本的范围更小。同时，CMFA方法在第1个特征向量上超出检测边界的浅色样本数均多于MFA方法，在图5中CMFA的浅色样本几乎都超出检测边界而MFA方法的浅色样本仍有一小部分位于边界范围中，在图6中，MFA方法位于检测边界外的浅色样本与检测边界内的浅色样本的比例接近1:1，而CMFA方法的比值接近2:1，很明显CMFA方法使两类样本的投影离的更远。综上CMFA方法在第1个投影向量上两类投影的分离程度优于MFA。

为了进一步验证方法的有效性，对两种规模的硬件木马分别进行10次检测实验，每次统计4种方法最优投影方向上位于检测边界外的样本数，如图7所示，横坐标为检测实验序号，纵坐标为位于检测边界外的最大投影数，虚线为检测边界，位于检测边界上方的点说明该次实验中该方法能检测出硬件木马。由图可知，核MMC和K-L变换均不能检测出两种硬件木马，其检出率为0%，对于硬件木马1而言，MFA和CMFA方法均能检测出硬件木马，检出率为100%，对于硬件木马2而言，MFA的检出率仅为20%，而CMFA的检出率仍为100%，具有更好的检测效果。

实验结果表明，CMFA方法具有比MFA，核MMC和K-L变换更好的检测效果，在确定好近邻参数的情况下，本文方法检测一次的平均时间为59.64 s，也优于K-L变换的100.64 s和核MMC的103.31 s。但是CMFA在选取近邻参数k时比较困难，需要进行多次实验，在一定程度上增大了时间复杂度，不过相对于芯片的安全性需求而言，CMFA方法的计算代价仍然可以接受。

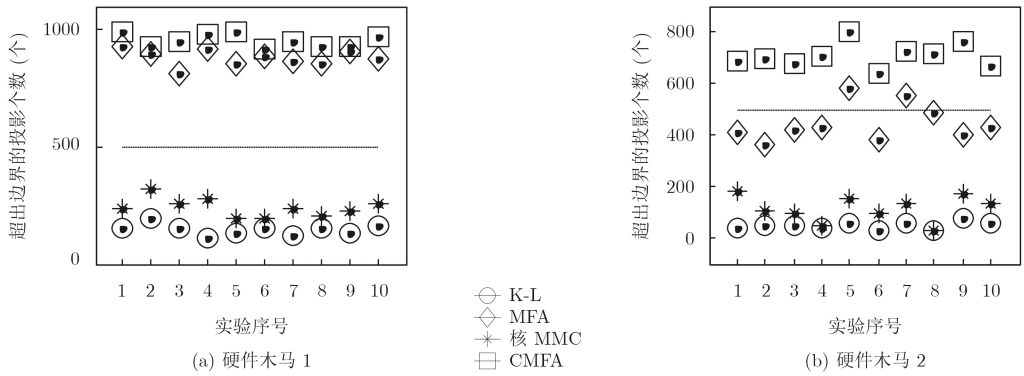


图 7 对两种硬件木马的10次检测结果

7　结束语

基于CMFA的硬件木马检测能够在不对数据分布作任何假设的前提下，有效提取“金片”和含硬件木马芯片之间旁路信号的特征差异，具有更高的检测精度，为硬件木马检测提供了一条新思路。为了进一步提高检测效果，可根据硬件木马可能植入电路中的位置研究相应的硬件木马激活策略，加速硬件木马检测。此外，还需采用遗传算法、蚁群算法等启发式算法或根据旁路信号统计模型寻找一种便捷确定最优近邻参数k的方法，提高CMFA方法的性能，实现硬件木马的快速检测。

参 考 文 献

[1] DOFE J, FREY J, and YU Qiaoyu. Hardware security assurance in emerging IoT applications[C]. 2016 IEEE International Symposium on Circuits and Systems, Montreal, Canada, 2016: 2050–2053. doi: 10.1109/ISCAS.2016.7538981.

[2] SUMATHI G, SRIVANI L, MURTHY D T, *et al*. A review on HT attacks in PLD and ASIC designs with potential defence solutions[J]. *IETE Technical Review*, 2018, 35(1): 64–77. doi: 10.1080/02564602.2016.1246385.

[3] CHAKRABORTY R S, WOLFF F, PAUL S, *et al*. *MERO*: A statistical approach for hardware Trojan detection[C]. The 11th International Workshop on Cryptographic Hardware and Embedded Systems, Switzerland, 2009: 396–410. doi: 10.1007/978-3-642-04138-9_28.

[4] SAHA S, CHAKRABORTY R S, NUTHAKKI S S, *et al*. Improved test pattern generation for hardware Trojan detection using genetic algorithm and Boolean satisfiability[C]. The 17th International Workshop on

Cryptographic Hardware and Embedded Systems, Saint-Malo, France, 2015: 577–596. doi: 10.1007/978-3-662-48324-4_29.

[5] LESPERANCE N, KULKARNI S, CHENG K T, *et al*. Hardware Trojan detection using exhaustive testing of k-bit subspaces[C]. The 20th Asia and South Pacific Design Automation Conference, Chiba, Japan, 2015: 755–760. doi: 10.1109/ASPDAC.2015.7059101.

[6] XUE Mingfu, HU Aiqun, and LI Guyue. Detecting hardware Trojan through heuristic partition and activity driven test pattern generation[C]. 2014 Communications Security Conference, Beijing, China, 2014: 1–6. doi: 10.1049/CP.2014.0728.

[7] AGRAWAL D, BAKTIR S, KARAKOYUNLU D, *et al*. Trojan detection using IC fingerprinting[C]. 2007 IEEE Symposium on Security and Privacy, Berkeley, USA, 2007: 296–310. doi: 10.1109/SP.2007.36.

[8] HE Jiaji, ZHAO Yiqiang, GUO Xiaolong, *et al*. Hardware Trojan detection through chip-free electromagnetic side-channel statistical analysis[J]. *IEEE Transactions on Very Large Scale Integration* (*VLSI*) *Systems*, 2017, 25(10): 2939–2948. doi: 10.1109/TVLSI.2017.2727985.

[9] XIAO Kan, ZHANG Xuehui, and TEHRANIPOOR M. A clock sweeping technique for detecting hardware trojans impacting circuits delay[J]. *IEEE Design & Test*, 2013, 30(2): 26–34. doi: 10.1109/MDAT.2013.2249555.

[10] 薛明富, 王箭, 胡爱群. 自适应优化的二元分类型硬件木马检测方法[J]. 计算机学报, 2018, 41(2): 439–451. doi: 10.11897/SP.J.1016.2018.00439.

XUE Mingfu, WANG Jian, and HU Aiqun. Adaptive optimization of two-class classification-based hardware Trojan detection method[J]. *Chinese Journal of Computers*, 2018, 41(2): 439–451. doi: 10.11897/SP.J.1016.2018.00439.

[11] 骆扬, 王亚楠. 物理型硬件木马失效机理及检测方法[J]. 物理学报, 2016, 65(11): 110602. doi: 10.7498/APS.65.110602.

LUO Yang and WANG Yanan. Physical hardware trojan failure analysis and detection method[J]. *Acta Physica Sinica*, 2016, 65(11): 110602. doi: 10.7498/APS.65.110602.

[12] 张鹏, 王新成, 周庆. 基于投影寻踪分析的芯片硬件木马检测[J]. 通信学报, 2013, 34(4): 122–126. doi: 10.3969/J.ISSN.1000-436x.2013.04.014.

ZHANG Peng, WANG Xincheng, and ZHOU Qing. Hardware Trojans detection based on projection pursuit[J]. *Journal on Communications*, 2013, 34(4): 122–126. doi: 10.3969/J.ISSN.1000-436x.2013.04.014.

[13] 李雄伟, 王晓晗, 张阳, 等. 一种基于核最大间距准则的硬件木马检测新方法[J]. 电子学报, 2017, 45(3): 656–661. doi: 10.3969/J.ISSN.0372-2112.2017.03.023.

LI Xiongwei, WANG Xiaohan, ZHANG Yang, *et al*. A new hardware Trojan detection method based on kernel maximum margin criterion[J]. *Acta Electronica Sinica*, 2017, 45(3): 656–661. doi: 10.3969/J.ISSN.0372-2112.2017.03.023.

[14] 赵毅强, 刘沈丰, 何家骥, 等. 基于自组织竞争神经网络的硬件木马检测方法[J]. 华中科技大学学报: 自然科学版, 2016, 44(2): 51–55. doi: 10.13245/J.HUST.160211.

ZHAO Yiqiang, LIU Shenfeng, HE Jiaji, *et al*. Hardware Trojan detection technology based on self-organizing competition neural network[J]. *Journal of Huazhong University of Science and Technology*: *Natural Science Edition*, 2016, 44(2): 51–55. doi: 10.13245/J.HUST.160211.

[15] YAN Shuicheng, XU Dong, ZHANG Benyu, *et al*. Graph embedding and extensions: A general framework for dimensionality reduction[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(1): 40–51. doi: 10.1109/TPAMI.2007.250598.

[16] 何进荣, 丁立新, 崔梦天, 等. 基于矩阵指数变换的边界Fisher分析[J]. 计算机学报, 2014, 37(10): 2196–2205.

HE Jinrong, DING Lixin, CUI Mengtian, *et al*. Marginal Fisher analysis based on matrix exponential transformation[J]. *Chinese Journal of Computers*, 2014, 37(10): 2196–2205.

[17] 李艳霞, 柴毅, 胡友强, 等. 不平衡数据分类方法综述[J]. 控制与决策, 2019, 34(4): 673–688. doi: 10.13195/J.KZYJC.2018.0865.

LI Yanxia, CHAI Yi, HU Youqiang, *et al*. Review of imbalanced data classification methods[J]. *Control and Decision*, 2019, 34(4): 673–688. doi: 10.13195/J.KZYJC.2018.0865.

王晓晗：男，1992年生，博士生，研究方向为芯片安全技术.

王 韬：男，1964年生，教授，博士，研究方向为信息安全与网络对抗.

李雄伟：男，1975年生，副教授，博士，研究方向为芯片安全技术.

张 阳：男，1984年生，讲师，博士，研究方向为集成电路安全.

黄长阳：男，1994年生，硕士生，研究方向为网络安全技术.