doi:10.11959/j.issn.1000-436x.2016282

基于图像来源分类的最小化虚警隐写分析模型

杨培韬，张卫明，俞能海
（中国科学技术大学中科院电磁空间信息重点实验室，安徽 合肥 230001）

摘 要：在实真场景中，在载体失配（CSM, cover source mismatch）条件下降低虚警率是隐写分析的一个巨大挑战，提出了一种新的模型来处理该问题。该方法由来源分类器首先判断图像的来源，继而利用相关来源图像训练而成的隐写分类器判断待测图像是否为载密。在这个过程中，通过对模型参数的调节减小虚警率。实验结果表明，这种方法可以在较大准确率的前提下最小化虚警率。

关键词：虚警率；失配；隐写分析；最小化虚警模型
中图分类号：TN309 **文献标识码**：A

Reducing false positives of steganalysis via classification of image-acquiring sources

YANG Pei-tao, ZHANG Wei-ming, YU Neng-hai
(CAS Key Laboratory of Electromagnetic Space Information, University of Science and Technology of China, Hefei 230001, China)

Abstract: In the real world, reducing false positive rates in the case of cover source mismatch (CSM) was a big challenge for steganalysis. A novel model was proposed to solve the problem. The proposed method determines the image-acquiring source firstly by a source detector and then detecting the steg images in each source with a steganalyzer trained for this source. The false positive rate was reduced by solving a parameter model. The experimental results show that this novel method can reach lower false positive rates for larger true positive rates.

Key words: false positive, mismatch, steganalysis, minimum false positive model

1 引言

隐写术是信息隐藏的一个分支[1]，可以将隐私数据嵌入到数字载体中。由于隐写前的载体对象与隐写后的载密对象难以区分，从而可以掩盖隐私数据的存在。正是因为这一特性，隐写术常常被极端分子用来从事犯罪活动。因此，与隐写术相对的隐写分析技术的发展显得格外重要。

隐写分析技术旨在检测隐私数据的存在[2]，传统的隐写分析模型是基于机器学习的理论设计的，将待测对象映射到某个特征空间，再通过二元分类器判断待测对象是否为载密。然而应用到真实场景中，传统的隐写分析模型将面临两大挑战，低虚警要求与失配现象。在真实场景中，载体对象的数量通常远远大于载密对象[3]。因此，传统隐写分析模型使用的分类器虚警率必须非常低，否则被误判为载密的载体对象会把系统淹没。另一方面，Fridrich等[4]指出传统隐写分析中训练集与测试集之间存在的各种失配，如训练集和测试集统计特征不一致导致的失配、嵌入率未知导致的失配、算法未知导致的失配等，会使传统隐写分析的错误率大幅提升。这种由失配导致的错误率提升足以说明传统的隐

收稿日期：2016-08-12；**修回日期**：2016-11-08
通信作者：张卫明，zhangwm@ustc.edu.cn
基金项目：国家自然科学基金资助项目（No.61572452, No.61502007, No.U1636201）；中国博士后科学基金资助项目（No.2015M582015）；中国科学院战略性先导专项基金资助项目（No.XDA06030601）
Foundation Items: The National Natural Science Foundation of China (No.61572452, No.61502007, No.U1636201), The China Postdoctoral Science Foundation (No.2015M582015), The Strategic Priority Research Program of the Chinese Academy of Sciences (No.XDA06030601)

写分析模型不适用于真实场景[5,6]。

目前，针对失配问题提出的隐写分析方法如下：Lubenko 等[7]认为利用简单分类器可以提高失配情况下的分类效果；在此基础上，Pasquet 等[8]引入了聚类的方法，提升了隐写分析的判别效果；此外，针对隐写算法的失配，文献[9,10]提出了基于迁移学习的隐写分析方法；针对量化表的失配，有基于特征映射变换的隐写分析方法。

这些工作成果都是基于传统的隐写分析方法，在一定程度上解决了失配隐写分析问题。然而，这些方法都没有考虑虚警率的要求。因此，针对这 2 个问题设计一套新的隐写分析系统有巨大的实际意义[11]。

本文以图像作为隐写分析的研究对象，以图像的生成设备不同作为失配问题的切入点，提出了最小化虚警模型（MFPM）。MFPM 的检测过程可以大致分为：1) 通过来源分类器判断测试图像的来源；2) 用该来源的图像训练而成的隐写分类器判断测试图像是否为载密；3) 通过参数的调节实现模型的虚警最小化。

2 传统隐写分析模型

一般而言，各种类型的数字媒体（如图像、视频、音频等）均可作为隐写术、隐写分析的研究对象，本文仅以图像为例，展示研究的效果。本文提出的模型依然适用于其他类型的数字媒体。

在检测载密图像过程中基于以下 2 个基本假设：

1) 使用的隐写算法已知；

2) 已知嵌入过程中的嵌入率。

传统的隐写分析方法检测流程描述如下。

1) 收集大量的载体图像，形成载体集，用 C 表示。基于上述假设，本文利用已知的隐写算法 $A(\cdot)$ 在固定的嵌入率下模拟隐私数据嵌入过程，从而生成载密图像集 S，这里 $S=A(C)$。为了方便表述，本文将 C 和 S 统称为训练集，用 T_r 表示。

2) 正如上文所述，隐写分析特征的提取操作用 $F_s(\cdot)$ 表示。通过将训练集的所有图像映射到特征空间，得到 $F_s(C)$ 和 $F_s(S)$。再选择合适的二元分类模型训练 $F_s(C)$ 和 $F_s(S)$，从而生成传统的隐写分类器 V。

3) 对于待测图像 x，首先计算 $F_s(x)$，再利用 V 检测 $F_s(x)$，判断 x 是否为载密。

3 最小化虚警模型

由于图像在拍摄过程中，拍摄设备会在图像中添加随机噪声和量化噪声，这些噪声会降低传统隐写分析模型的检测准确性。因此，在最小化虚警模型中增加了对图像来源判断的处理。具体如下所述。

1) 与传统隐写分析流程相同，首先收集大量的载体图像。不同的是本文按照图像来源的不同将载体图像分为若干个载体子集，记为 C_i，$i=1,2,\cdots,N$，其中，N 表示载体子集的总数。载体子集中的图像均来自相同型号的图像采集设备。

2) 相似地，本文利用来源特征 $F_f(\cdot)$ 实现图像的来源判断。将所有载体子集中的图像映射到特征空间得到 $F_f(C_i)$，$i=1,2,\cdots,N$。由于 $N\geqslant 2$，本文选择多元分类模型训练 $F_f(C_i)$，生成来源分类器 V_f。

3) 同样地，本文利用特征 $F_s(\cdot)$ 判断待测图像是否为载密。首先模拟生成载密子集 S_i（$S_i=A(C_i)$），再分别将所有载体与载密子集中的图像映射到特征空间中，生成 $F_s(C_i)$ 和 $F_s(S_i)$，$i=1,2,\cdots,N$。

4) 训练 $F_s(C_i)$ 和 $F_s(S_i)$，得到第 i 组隐写分类器 V_i。循环此操作，最终生成 N 个隐写分类器。

以上为 MFPM 的训练过程，而其测试过程如图 1 所示。

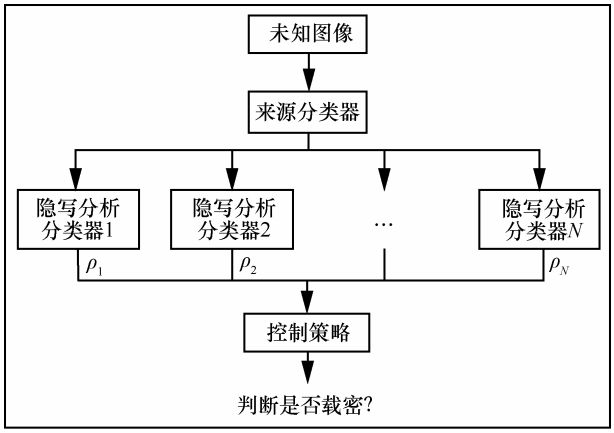


图 1 MFPM 的测试过程

1) 对于待测图像 y，首先计算 $F_f(y)$，通过 V_f 判断 y 的图像来源。这里不失一般性，本文假设 y 来自第 k 组图像来源。

2) 计算 $F_s(y)$。由于 y 来自第 k 组图像，本文用 V_k 对 y 进行载体载密判断。V_k 的输出为

$$V_k(y)=\begin{cases}0, & y\text{为载体图像}\\ 1, & y\text{为载密图像}\end{cases} \tag{1}$$

每次测试有且仅有一个隐写分类器 V_k 会对 y

进行载体载密判断。对于不响应的隐写分类器，设置 $V_i=0$，$i=1,2,\cdots,k-1,k+1,\cdots,N$。

3) 本文利用函数 F_{cs} 生成最终判决结果为

$$F_{\mathrm{cs}}(\rho_1,\cdots,\rho_N;V_1,\cdots,V_N)=R\left\{\sum_{i=1}^{N}\rho_i V_i\right\} \tag{2}$$

其中，ρ_i 是调整参数，且 $0\leqslant\rho_i\leqslant 1$，$R\{p\}$ 为生成随机数操作（以概率 p 生成 1，以概率 $1-p$ 生成 0）。最终的结果满足

$$F_{\mathrm{cs}}=\begin{cases}0, & y\text{为载体图像}\\ 1, & y\text{为载密图像}\end{cases} \tag{3}$$

其中，式(2)中的参数 $\rho_i\,(i=1,2,\cdots,N)$ 为最小化模型的虚警率。为了简化处理过程，假设本文可以准确地识别未知图像的来源。对于传统隐写分析模型，V_i 的检错率 P_{Ei} 满足

$$P_{Ei}=\frac{1}{2}(P_{FPi}+P_{FNi}) \tag{4}$$

其中，P_{FPi} 与 P_{FNi} 分别表示 V_i 的虚警率与漏警率。由于 MFPM 受到参数 ρ_i 的影响，此时 V_i 的检错率 P_i 为

$$\begin{aligned}P_i &=\frac{1}{2}\left\{\rho_i P_{FPi}+\left[1-\rho_i(1-P_{FNi})\right]\right\}\\ &=\frac{1}{2}(1-\rho_i)+\rho_i P_{Ei}\end{aligned} \tag{5}$$

由于 $0\leqslant\rho_i\leqslant 1$，$0\leqslant P_{Ei}\leqslant\frac{1}{2}$，在引入参数 ρ_i 后，有 $P_i\geqslant P_{Ei}$，即提高检错率；但是另一方面有 $\rho_i P_{FPi}\leqslant P_{FPi}$，即降低虚警率。因此，本文可以通过调整 $\rho_i\,(i=1,2,\cdots,N)$ 实现模型的最小虚警为

$$\begin{aligned}&\min\frac{1}{N}\sum_{i=1}^{N}\rho_i P_{FPi}\\ &\text{s.t.}\quad 0\leqslant\frac{1}{N}\sum_{i=1}^{N}P_i\leqslant P_{\mathrm{th}}\end{aligned} \tag{6}$$

其中，P_{th} 是模型检错率的上限。P_{FPi} 与 P_{FNi} 可以通过实验得到，因此，式(6)只要给定模型检错率的上限 P_{th}，即可计算出使模型达到最小虚警的参数 $\rho_i\,(i=1,2,\cdots,N)$。

4　实施方法

4.1　实验对象选择

本文以空域图像为例验证上述方法。由手机、相机等设备直接拍摄的图像为 JPEG 格式，所以本文在真实场景中用于隐写的空域图像大多是由 JPEG 格式的图像解压缩得到的。因此，本文采用 JPEG 解压缩空域图像作为最小化虚警模型的实验对象。

4.2　来源分类方法选择

由第 3 节可知，能否准确判断图像的来源与整个最小化虚警模型的检测准确率的高低有着直接的关联。本文选择 Fridrich 等提出的 PCE[12~15]（peak-to-correlation energy ratio）为 $F_{\mathrm{f}}(\cdot)$。

类似于归一化的相关系数，PCE 通常被用来计算 2 个离散信号间的相似度。由于图像与错误来源计算得出的 PCE 远小于图像与正确来源计算的 PCE 的值，因此，PCE 通常用来判断图像的来源。本文先利用快速离散傅里叶变换计算图像与来源间的互相关，再通过互相关计算 PCE 的值。

4.3　隐写分析方法选择

本文采用的隐写算法（上文提到的 $A(\cdot)$）为非自适应的隐写算法（LSB matching）。根据第 2 节的假设 2)，本文的实验采用的嵌入率为 0.05、0.1、0.2 以及 0.4 bit/pixel。另外，选用的隐写分类器为 ensemble 分类器（版本为 2.0，默认设置，下载地址为 http://dde.binghamton.edu/download/ ensemble/）[16]。采用的隐写分析特征为 34 671 维度的 SRM（spatial rich model）特征[17]。

SRM 首先计算 22 个一阶及三阶残差矩阵、12 个二阶残差矩阵、2 个 SQUARE 残差矩阵、10 个 EDGE 3×3 及 EDGE 5×5 残差矩阵，共计 22+22+12+2+10+10=78 个残差矩阵。分别计算上述残差矩阵的四阶马尔可夫特征，范围参数 T=2，即每个残差矩阵有 $(2T+1)^4=625$ 维。利用符号对称性及方向对称性降低残差矩阵个数及特征维度。可将一阶及三阶残差矩阵降至 12 个、二阶残差矩阵降至 7 个、SQUARE 残差矩阵降至 2 个、EDGE 3×3 及 EDGE 5×5 残差矩阵降至 6 个。可将 12 个一阶特征降至 169 维、其他 33 个特征降至 325 维，共计 $12\times169+33\times325=12\,753$ 维。上述 12 753 维度特征采用步长 $q=1$ 进行量化，若量化步长按照式(7)确定，则可以得到 $2\times(2\times169+10\times325)+3\times(10\times169+23\times325)=34\,671$ 维的 SRM 特征，其中，c 为残差矩阵阶数。

$$q\in\begin{cases}\{c,1.5c,2c\}, & c>1\\ \{1,2\}, & c=1\end{cases} \tag{7}$$

5　实验分析

5.1　图像库准备

由 4.1 节的论述，为了保证实验的准确性，本

文收集了由 200 多种不同型号的设备拍摄的 70 000 余张 JPEG 格式图像。根据在实验中对各来源的图像有数量和质量（主要指的是图像清晰度）上的要求，本文最终选择了 5 种来源的 13 601 张图像作为本文实验的原始图像。这 5 种图像分别来源于 iPhone 4s、iPhone 5、Nikon D3100、Nikon D700 和 Sony TX1，依次记为 iP4、iP5、NK3、NK7 和 ST。

通过软件（imageMagick）将这些原始图像解压缩成 24 位 TIFF 格式的彩色图像。为了进一步增加实验图像数量，本文将解压得到的彩色图像裁剪成 1 024×1 024 像素的图像块，再将各图像块采样至 512×512，最后将这些 512×512 的图像块转化成 PNG 格式 8 位的灰度图像。通过上述方法，本文共计得到 41 556 张灰度图像，这些灰度图像即为本文实验的图像库，具体参见表 1。

表 1 各图像来源的图像数量

图像来源	图像数量/张
iP4	7 136
iP5	10 000
NK7	8 420
NK3	10 000
ST	6 000
合计	41 556

5.2 失配现象验证

Fridrich 等[4]已通过实验表明失配现象会对传统隐写分析的结果造成不利影响，本文用表 1 中的图像重现该实验。

本文将每个来源的图像分为 2 个集合：由随机选取的 1 500 张图像组成的测试集和由剩下的图像组成的训练集（合计得到了 5 个测试集与 5 个训练集）。在失配实验中，本文随机选取一个训练集中 4 000 张图像用来训练 V_i，再利用 V_i ($i=1,2,\cdots,5$)依次检测 5 个测试集中的图像。表 2 和表 3 中总结了在 0.1 bit/pixel 和 0.4 bit/pixel 嵌入率下失配现象对传统隐写分析模型的影响。表 2 和表 3 中的数值为隐写分类器检测的错误率。本文在表 2 和表 3 中用加粗的方式标记了在训练集与测试集匹配情况下的测试结果。

通过表 2 和表 3，可以看出如下特点。

1) 表 2 和表 3 中的检错率明显大于文献[12,18,19]中的数值，这是由于本文采用的图像库是由 JPEG 图像解压缩得到的（存在量化因素），并且本文采用 imageMagick 软件而非 Matlab 对图像进行格式转换。

表 2 嵌入率在 0.1 bit/pixel 条件下的检错率

训练集	测试集				
	iP4	iP5	NK7	NK3	ST
iP4	**0.322 8**	0.346 2	0.488 3	0.401 2	0.481 2
iP5	0.444 7	**0.265 2**	0.486 1	0.389 6	0.461 6
NK7	0.439 5	0.426 2	**0.206 7**	0.430 5	0.486 2
NK3	0.439 0	0.378 4	0.420 3	**0.272 7**	0.477 4
ST	0.490 8	0.475 5	0.492 7	0.440 8	**0.387 7**

表 3 嵌入率在 0.4 bit/pixel 条件下的检错率

训练集	测试集				
	iP4	iP5	NK7	NK3	ST
iP4	**0.118 5**	0.159 1	0.244 0	0.188 8	0.472 5
iP5	0.417 7	**0.091 2**	0.274 1	0.287 1	0.378 2
NK7	0.285 3	0.245 1	**0.025 1**	0.204 9	0.414 1
NK3	0.352 9	0.222 7	0.237 4	**0.069 1**	0.425 6
ST	0.480 0	0.396 4	0.451 1	0.368 7	**0.212 7**

2) 尽管图像不一致，还是清晰展示了失配现象造成的影响：失配情况下的检错率远远大于匹配情况下的数值，并且对于同一图像来源而言，随着嵌入率的提高，失配现象造成的影响逐步提高。另外，在同一嵌入率下，失配现象对不同的训练集图像造成的影响完全不同。

5.3 图像来源测试

按照 5.2 节中提到的方法，本文将图像集分成 5 个测试集与 5 个训练集。利用训练集中的全部图像训练 V_{f}，并用 V_{f} 判断 5 个测试集中全部图像的来源，测试结果如图 2 所示。来源分类器的平均检测正确率为 87.04%，甚至部分图像来源（如 NK3、ST）的检测准确率近乎 100%。

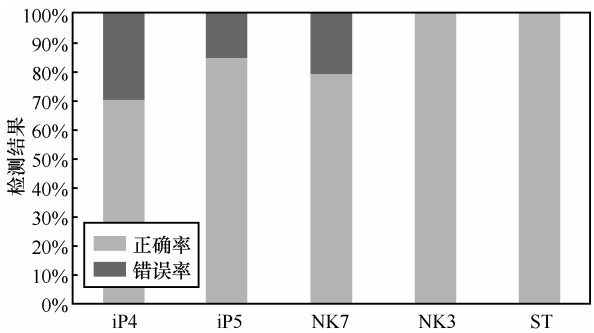


图 2 来源分类器的检测结果

5.4　综合实验

为了更好地说明实验结果，首先用 5.1 节的数据库进行传统隐写分析实验。考虑到图像的数量过大（共计 41 556 张），本文在每个图像来源中随机选择了 1 500 张（共计 7 500 张）。与第 2 节描述的方法相同，本文从中随机选择 4 000 张图像作为训练集，剩下的 3 500 张图像作为测试集，利用 SRM 特征进行传统隐写分析检测。

另一方面，按照 5.2 节的方法将所有的图像分成 5 个训练集与 5 个测试集。在根据第 3 节所描述的方法获得 V_{f} 和 V_i （ $i=1,2,\cdots,5$ ）。最后按照第 3 节描述的测试流程测试 5 个测试集中的 7 500 张图像。这里本文先设置 ρ_i=1， $i=1,2,\cdots,N$ 。

图 3 所示为传统隐写分析模型与最小化虚警模型的检测结果，可以看出 MFPM 的检错率始终低于传统隐写分析模型。所以，MFPM 在失配情况下有利于提升隐写分析的检测效果，但改进并不明显。不过，本文提出 MFPM 模型的重点在于控制虚警率。

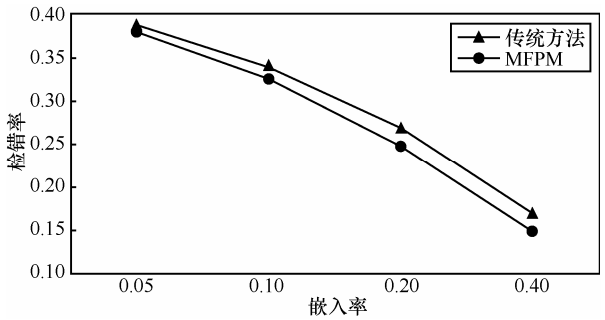
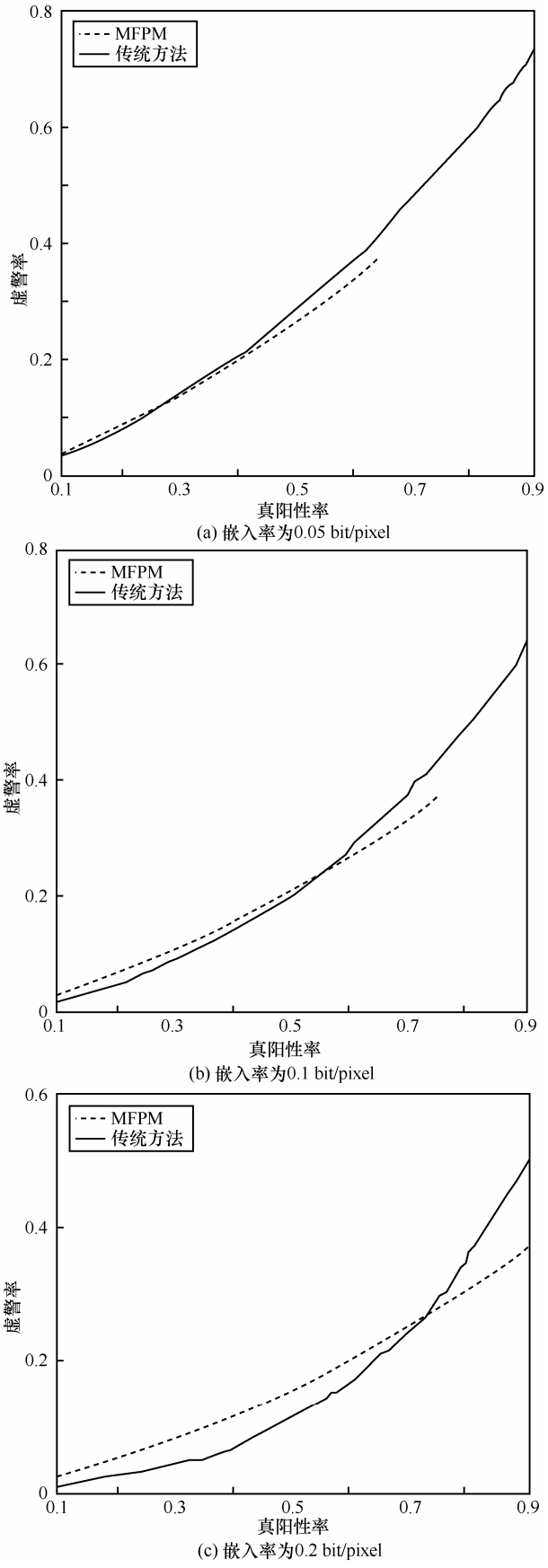


图 3　2 种模型的测试结果比较

5.5　参数调节

第 3 节提到，本文可以通过参数（ ρ_i ）的调节控制检测结果的虚警率。本文重做 5.4 节的实验，此次聚焦于 2 种模型的虚警率，实验结果如图 4 所示。

由于在式(6)中存在检错率上限 P_{th}，因此 MFPM 中的准确率存在上限。另外可以发现，MFPM 的虚警率呈线性增加，而传统隐写分析模型的虚警率呈指数增加。这导致当准确率较大时，MFPM 的虚警率远小于传统隐写分析模型。MFPM 是针对真实场景设计的隐写分析模型，而在真实场景中，本文尽可能准确地识别载密对象。因此，MFPM 的高准确率、低虚警率特性正是真实场景的隐写分析所需要的[20]。



(a) 嵌入率为0.05 bit/pixel

(b) 嵌入率为0.1 bit/pixel

(c) 嵌入率为0.2 bit/pixel

图 4　2 种模型的虚警率比较

6 结束语

本文以图像来源不同造成的失配现象为切入点，提出了最小化虚警模型。与传统隐写分析模型相比，MFPM 可以通过对参数的调整降低虚警率。本文在参数计算的过程中假设来源分类的结果是准确无误的，而根据 5.3 节的实验结果可知，尽管来源分类的准确率很高，但依然存在误判。因此，在接下来的工作中，本文在参数调整的过程中充分考虑来源分类的误判造成的影响。

另一方面，本文所提出的最小化虚警模型是一个一般性的模型。仅以解压缩的 JPEG 图像为例验证 MFPM 的可行性，当然 MPFM 也适用于其他的失配场景以及其他类型的载体。

参考文献：

[1] GOLJAN M, FRIDRICH J, CHEN M. Sensor noise camera identification: countering counter-forensics[C]//SPIE Media Forensics and Security II. 2010:75410S.

[2] FRIDRICH J. Steganography in digital media: principles, algorithms, and applications[M]. Cambridge University Press, 2009.

[3] PEVNY T, KER A D. Towards dependable steganalysis[C]//SPIE Media Watermarking, Security, and Forensics. 2015:94090I.

[4] KODOVSKY J, SEDIGHI V, FRIDRICH J. Study of cover source mismatch in steganalysis and ways to mitigate its impact[C]// SPIE Media Watermarking, Security, and Forensics. 2014:90280J.

[5] BARNI M, CANCELLI G, ESPOSITO A. Forensics aided steganalysis of heterogeneous images[C]//IEEE Conference Acoustics Speech and Signal Process, 2010: 1690-1693.

[6] CANCELLI G, DOERR G, BARNI M. A comparative study of ±1 steganalyzers[C]//IEEE Multimedia Signal Process. Workshop, 2008: 791-794.

[7] LUBENKO I, KER A D. Steganalysis with mismatched cover: do simple classifiers help[C]//ACM Workshop on Multimedia and Security. 2012: 11-18.

[8] PASQUET J, BRINGAY S, CHAUMONT M. Steganalysis with cover-source mismatch and a small learning database[C]//22nd European Signal Processing Conference (EUSIPCO). IEEE, 2014: 2425-2429.

[9] LI X, KONG X, WANG B. Generalized transfer component analysis for mismatched JPEG steganalysis[C]//In IEEE International Conference on Image Processing, 2013: 4432-4436.

[10] ZENG L, KONG X, LI M. JPEG quantization table mismatched steganalysis via robust discriminative feature transformation[C]//In SPIE/IS&T Electronic Imaging. International Society for Optics and Photonics, 2015: 94090U.

[11] KER A D, BAS P, FRIDRICH J. Moving steganography and steganalysis from the laboratory into the real world[C]//The 1st ACM Workshop on Information Hiding and Multimedia Security, 2013: 45-58.

[12] FRIDRICH J, GOLJAN M. Determining approximate age of digital images using sensor defects[C]//SPIE Media Watermarking, Security, and Forensics III, 2011: 788006.

[13] GOLJAN M, FRIDRICH J, FILLER T. Managing a large database of camera fingerprints[C]//SPIE Media Forensics and Security II, 2010: 754108.

[14] GOLJAN M, FRIDRICH J, CHEN M. Defending against fingerprint-copy attack in sensor-based camera identification[J]. In IEEE Transactions on Information Security and Forensics. 2010, 6(1): 227-236.

[15] GOLJAN M, FRIDRICH J. Sensor-fingerprint based identification of images corrected for lens distortion[C]//SPIE Media Watermarking, Security, and Forensics. 2012: 83030H.

[16] KODOVSKY J, FRIDRICH J, HOLUB V. Ensemble classifiers for steganalysis of digital media[J]. In IEEE Transaction on Information Forensics and Security, 2012, 7(2): 432-444.

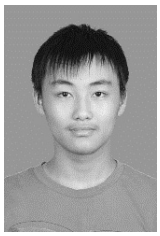
[17] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images[J]. In IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.

[18] COGRANNE R, ZITZMANN C, RETRAINT F. Statistical detection of LSB matching using hypothesis testing theory[C]//The 14th International Conference on Information Hiding. 2013: 46-62.

[19] DENEMARK T, FRIDRICH J. Detection of content adaptive LSB matching: a game theory approach[C]//SPIE Media Watermarking, Security, and Forensics. 2014: 902804.

[20] 李凤华, 殷丽华, 吴巍, 等. 天地一体化信息网络安全保障技术研究进展及发展趋势[J]. 通信学报, 2016, 37(11): 156-166.
LI F H, YIN L H, WU W, et al. Research status and development trends of security assurance for space-ground integration information network[J]. Journal on Communications, 2016, 37(11): 156-166.

作者简介：



杨培韬（1991-），男，安徽安庆人，中国科学技术大学硕士生，主要研究方向为隐写分析。

张卫明（1976-），男，河北保定人，博士，中国科学技术大学副教授，主要研究方向为信息隐藏、密码学。

俞能海（1964-），男，安徽无为人，博士，中国科学技术大学教授，主要研究方向为视频处理与多媒体通信、无线通信中的信号处理与分析、信息隐藏与信息安全。