$$\Theta=\{\Delta,Ms,C_{Ms},U,q^U\}$$

采取的单分识别方案按定义 9 描述为

$$U=(SPO,fea,RF),$$

其中：

(1) SPO 为方案工作流程，因其与 $SLRO$（第 2.2 节）基本相同，不再具体描述；

(2) fea 为方案对密文数据提取的单分特征，可为 3.4.2 节中任意一种；

(3) RF 表示方案选用随机森林分类算法作为识别算法.

因识别算法的选用，称该方案为基于随机森林的密码体制单分识别方案.

4.3　分层识别方案

现在提出基于随机森林的密码体制分层识别方案.

在密码体制识别问题 $\Gamma=(M,J,h^J)$ 中，基于随机森林的密码体制分层识别方案按定义 2 描述为 $J=(TLSO,FEA,RF)$，其中：

(1) $TLSO$ 为方案的工作流程；

(2) FEA 为方案对密文数据提取的特征；

(3) RF 表示识别中采取随机森林分类算法.

方案工作流程 $TLSO$ 由识别情境、训练阶段和测试阶段组成，下面以过程形式对其描述（图 2 为相应流程图）.

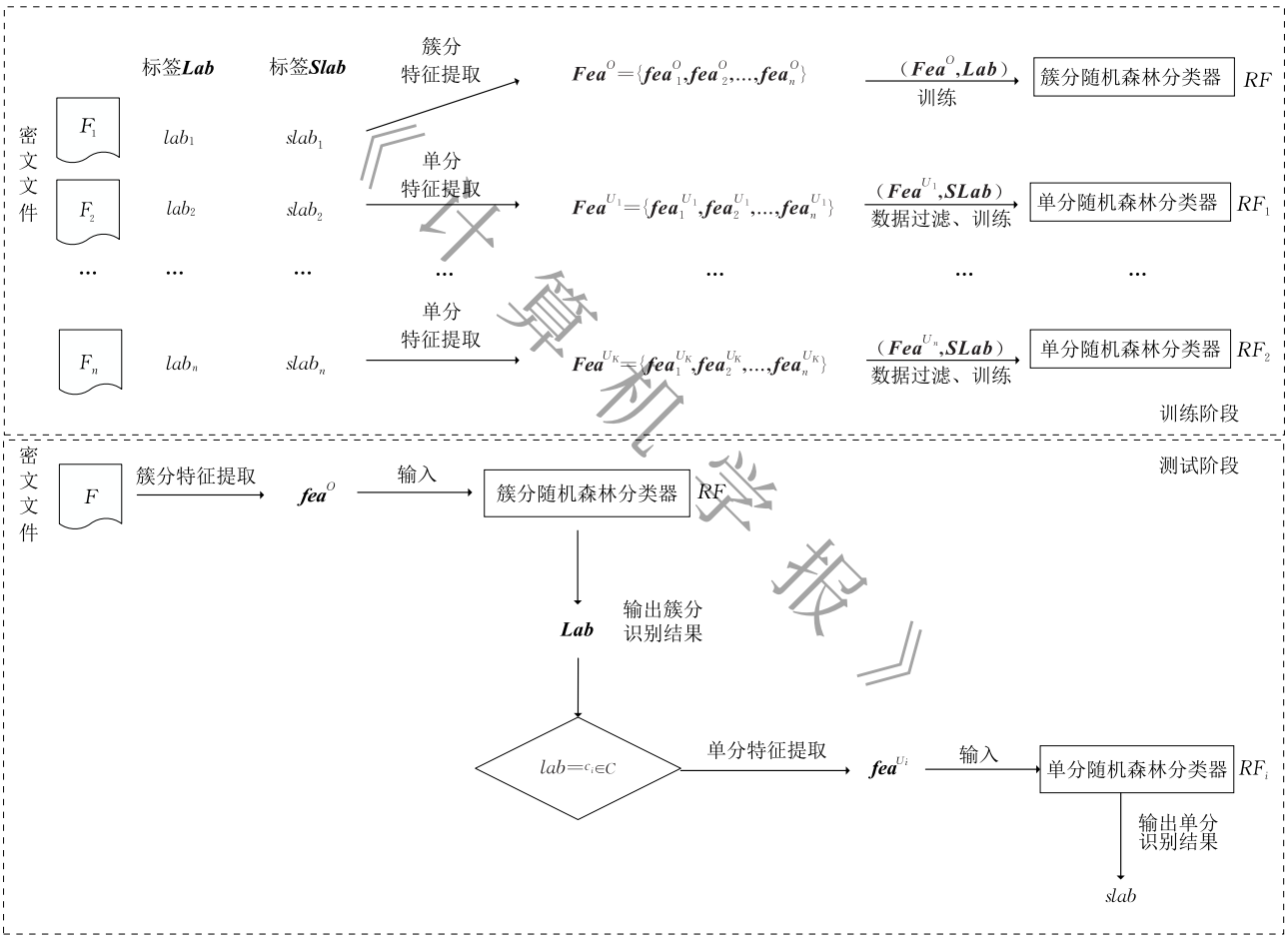

图 2　密码体制分层识别方案流程图

过程 3.　$TLSO$.

识别情境

识别情境由簇分设置

$$\Delta=(M,C,f,O,p^O)$$

和 K 个单分设置

$$\Theta_1=(\Delta,Ms_1,c_{Ms_1},U_1,q^{U_1}),$$
$$\Theta_2=(\Delta,Ms_2,c_{Ms_2},U_2,q^{U_2}),$$
$$\cdots,$$
$$\Theta_K=(\Delta,Ms_K,c_{Ms_K},U_K,q^{U_K})$$

组成，简记为 $S=(\Delta,K,\Theta_1,\cdots,\Theta_K)$.

训练阶段

训练阶段包含如下步骤：

(1) 采集已知簇类别和其对应密码体制的一组密文文件 $F_1,F_2,\cdots,F_n$，其中 n 为文件个数；

(2) 对密文文件内容数据进行特征提取，得簇分特征集

$$\boldsymbol{Fea}^O=\{\boldsymbol{fea}_1^O,\boldsymbol{fea}_2^O,\cdots,\boldsymbol{fea}_n^O\}$$

和 K 组单分特征集

$$\boldsymbol{Fea}^{U_1}=\{\boldsymbol{fea}_1^{U_1},\boldsymbol{fea}_2^{U_1},\cdots,\boldsymbol{fea}_n^{U_1}\},$$
$$\boldsymbol{Fea}^{U_2}=\{\boldsymbol{fea}_1^{U_2},\boldsymbol{fea}_2^{U_2},\cdots,\boldsymbol{fea}_n^{U_2}\},$$
$$\cdots,$$
$$\boldsymbol{Fea}^{U_K}=\{\boldsymbol{fea}_1^{U_K},\boldsymbol{fea}_2^{U_K},\cdots,\boldsymbol{fea}_n^{U_K}\};$$

(3)记 n 个文件的簇类别(已知)为一个 n 维向量 $\boldsymbol{Lab}=(lab_1,lab_2,\cdots,lab_n)$,称二元组($\boldsymbol{Fea}^O$,$\boldsymbol{Lab}$)为带标签的簇分数据;

(4)将带标签的簇分数据($\boldsymbol{Fea}^O$,$\boldsymbol{Lab}$)提交给随机森林分类算法,进行簇分分类模型的训练;

(5)记 n 个文件所属密码体制的标签组成为一个 n 维向量

$$\boldsymbol{SLab}=(slab_1,slab_2,\cdots,slab_n),$$

称二元组($\boldsymbol{Fea}^{U_i}$,$\boldsymbol{SLab}$),$i=1,2,\cdots,K$ 为带标签的单分数据;

(6)在单分设置 $\Theta_i=(\Delta,Ms_i,c_{Ms_i},U_i,q^{U_i})$ 下,将带标签的单分数据

($\boldsymbol{Fea}^{U_i}$,$\boldsymbol{SLab}$),$i=1,2,\cdots,K$

进行过滤,剔除掉不属于密码体制子集 Ms_i 的数据,并把过滤后的数据提交给随机森林分类算法,进行单分设置 Θ_i 下分类模型的训练.

测试阶段

测试阶段包含如下步骤:

(1)对一个待识别密文文件 F 的内容数据进行特征提取,得到簇分特征 $\boldsymbol{fea}^O$;

(2)将簇分特征 $\boldsymbol{fea}^O$ 输入到在训练阶段训练好的簇分分类模型中,后者给出簇分识别结果,即簇类别 lab;

(3)若 $lab=c_i\in C$,对密文文件 F 的内容数据进行特征提取,得到单分特征 $\boldsymbol{fea}^{U_i}$;

(4)将单分特征 $\boldsymbol{fea}^{U_i}$ 输入到在训练阶段训练好的单分设置 Θ_i 下的分类模型中,后者给出单分识别结果,即密码体制标签 $slab$.

根据过程 3 对方案工作流程 $TLSO$ 的介绍,在分层识别方案 $J=(TLSO,FEA,RF)$ 中,提取的特征 FEA 即为上述簇分特征与单分特征的组合.

5 实验与分析

5.1 密文数据采集

为开展实验,本文采集了多种密码体制的产生的密文数据.具体地,考察 13 种密码体制:Substitution、Permutation、Trivium、Sosemanuk、Grain、RC4、AES、Camellia、DES、3DES、SMS4、RSA、ECC.在此基础上,为更好地在簇分识别中考察密码体制的共性和差异、在单分识别中揭示密码体制不同参数设置对密文特性的影响,通过引入多样化的参数设置,将它们扩充为 42 种密码体制,分别为:

(1)选定密钥的 Substitution,记为 Sub;

(2)选定密钥的 Permutation,记为 Perm;

(3)~(6)选定 4 个不同密钥的 Trivium,分别记为 T1、T2、T3、T4;

(7)~(10)选定 4 个不同密钥的 Sosemanuk,分别记为 So1、So2、So3、So4;

(11)~(14)选定 4 个不同密钥的 Grain,分别记为 G1、G2、G3、G4;

(15)~(18)选定 4 个不同密钥的 RC4,分别记为 rc4-1、rc4-2、rc4-3、rc4-4;

(19)~(24)对于不同工作模式、不同分组长度的 AES-128-ecb、AES-128-cbc、AES-192-ecb、AES-192-cbc、AES-256-ecb、AES-256-cbc,分别选定密钥,记为 A1、A2、A3、A4、A5、A6;

(25)~(30)对于不同工作模式、不同分组长度的 Camellia-128-ecb、Camellia-128-cbc、Camellia-192-ecb、Camellia-192-cbc、Camellia-256-ecb、Camellia-256-cbc,分别选定密钥记为 C1、C2、C3、C4、C5、C6;

(31)~(32)对于不同工作模式的 DES-ecb、DES-cbc,分别选定密钥,记为 D1、D1-2;

(33)~(34)对于不同工作模式的 3DES-ecb、3DES-cbc,分别选定密钥,记为 D3-1、D3-2;

(35)~(38)对于不同工作模式的 SMS4-ecb、SMS4-cbc、SMS4-cfb、SMS4-ofb,分别选定密钥,记为 S1、S2、S3、S4;

(39)~(40)对于不同参数规模的 RSA-1024、RSA-2048,分别选定密钥,记为 R1、R2;

(41)~(42)对于不同参数规模的 ECC-224、ECC-384,分别选定密钥,记为 E1、E2.

在数据采集阶段,从 1000 个随机生成的大小固定为 512 KB 的明文文件出发,对上述每一密码体制,以这 1000 个明文文件为输入,加密得到 1000 个密文文件(特别地,对两种 ECC 体制,各得 500 个密文文件),并对这些密文文件截断,将文件长度仍控制为 512 KB.共计得到 41000 个 512 KB 的密文文件.

除 C、Java 平台程序实现外,密码体制的实现使用了开源工具 OpenSSL① 及 GmSSL②.各密码体制的具体设置及实现方式,参见表 2.

① OpenSSL—Cryptography and SSL/TLS Toolkit. https://www.openssl.org/ 2016,11,10

② GmSSL—支持SM2,SM3,SM4算法的OpenSSL分支. http://gmssl.org/ 2016.5.12

表 2　进行密文数据采集的 42 种密码体制

记号	密钥	工作模式	分组长度	参数规模	实现方式	全称	记号	密钥	工作模式	分组长度	参数规模	实现方式	全称
Sub	选定	—	—	密钥	C	Substitution	A4	选定	cbc	192	固定	OpenSSL	AES-192-cbc
Perm	选定	—	—	密钥	C	Permutation	A5	选定	ecb	256	固定	OpenSSL	AES-256-ecb
T1	选定	—	—	固定	Java	Trivium-Key1	A6	选定	cbc	256	固定	OpenSSL	AES-256-cbc
T2	选定	—	—	固定	Java	Trivium-Key2	C1	选定	ecb	128	固定	OpenSSL	Camellia-128-ecb
T3	选定	—	—	固定	Java	Trivium-Key3	C2	选定	cbc	128	固定	OpenSSL	Camellia-128-cbc
T4	选定	—	—	固定	Java	Trivium-Key4	C3	选定	ecb	192	固定	OpenSSL	Camellia-192-ecb
So1	选定	—	—	固定	Java	Sosemanuk-Key1	C4	选定	cbc	192	固定	OpenSSL	Camellia-192-cbc
So2	选定	—	—	固定	Java	Sosemanuk-Key2	C5	选定	ecb	256	固定	OpenSSL	Camellia-256-ecb
So3	选定	—	—	固定	Java	Sosemanuk-Key3	C6	选定	cbc	256	固定	OpenSSL	Camellia-256-cbc
So4	选定	—	—	固定	Java	Sosemanuk-Key4	D1	选定	ecb	56	固定	OpenSSL	DES-ecb
G1	选定	—	—	固定	Java	Grain-Key1	D1-2	选定	cbc	56	固定	OpenSSL	DES-cbc
G2	选定	—	—	固定	Java	Grain-Key2	D3-1	选定	ecb	56	固定	OpenSSL	3DES-ecb
G3	选定	—	—	固定	Java	Grain-Key3	D3-2	选定	cbc	56	固定	OpenSSL	3DES-cbc
G4	选定	—	—	固定	Java	Grain-Key4	S1	选定	ecb	128	固定	GmSSL	SMS4-ecb
rc4-1	选定	—	—	固定	Java	RC4-Key1	S2	选定	cbc	128	固定	GmSSL	SMS4-cbc
rc4-2	选定	—	—	固定	Java	RC4-Key2	S3	选定	cfb	128	固定	GmSSL	SMS4-cfb
rc4-3	选定	—	—	固定	Java	RC4-Key3	S4	选定	ofb	128	固定	GmSSL	SMS4-ofb
rc4-4	选定	—	—	固定	Java	RC4-Key4	R1	选定	—	—	1024	Java	RSA1024
A1	选定	ecb	128	固定	OpenSSL	AES-128-ecb	R2	选定	—	—	2048	Java	RSA2048
A2	选定	cbc	128	固定	OpenSSL	AES-128-cbc	E1	选定	—	—	224	Java	ECC224
A3	选定	ecb	192	固定	OpenSSL	AES-192-ecb	E2	选定	—	—	384	Java	ECC384

5.2　簇分可行性探讨

由 3.2 节，首先通过实验探讨 CM-簇分 $\Delta_{CM}=(M,C_{CM},f_{CM},O_{CM},p_{CM}^{O})$、$CSN$-簇分 $\Delta_{CSN}=(M,C_{CSN},f_{CSN},O_{CSN},p_{CSN}^{O})$、$CSBP$-簇分 $\Delta_{CSBP}=(M,C_{CSBP},f_{CSBP},O_{CSBP},p_{CSBP}^{O})$ 在相应识别方案

$$O_{CM}=(CPO,fea_{CM},RF)、$$
$$O_{CSN}=(CPO,fea_{CSN},RF)、$$
$$O_{CSBP}=(CPO,fea_{CSBP},RF)$$

下的可行性.

在所给簇分设置 Δ_{CM}、Δ_{CSN}、Δ_{CSBP} 下，将密文数据采集阶段收集到的所有密文数据纳入考虑中，即令 Δ_{CM}、Δ_{CSN}、Δ_{CSBP} 中密码体制集合

$$M=\left\{\begin{array}{l}\text{Sub},\text{Perm},\\ \text{T1},\text{T2},\text{T3},\text{T4},\\ \text{So1},\text{So2},\text{So3},\text{So4},\\ \text{G1},\text{G2},\text{G3},\text{G4},\\ \text{rc4-1},\text{rc4-2},\text{rc4-3},\text{rc4-4},\\ \text{A1},\text{A2},\text{A3},\text{A4},\text{A5},\text{A6},\\ \text{C1},\text{C2},\text{C3},\text{C4},\text{C5},\text{C6},\\ \text{D1-1},\text{D1-2},\text{D3-1},\text{D3-2},\\ \text{S1},\text{S2},\text{S3},\text{S4},\\ \text{R1},\text{R2},\text{E1},\text{E2}\end{array}\right\}.$$

由于在定义 5～7 中已明确给出簇集合 C_{CM}、C_{CSN}、C_{CSBP}，相应地，可在当前设定 M 的基础上将簇分映射 f_{CM}、f_{CSN}、f_{CSBP} 具体化：

令 $M[i]$ 表示集合 M 的第 i 个元素，则

$$f_{CM}(M[i])=\begin{cases}c_C, & 1\le i\le 2\\ c_M, & 3\le i\le 42\end{cases},$$

$$f_{CSN}(M[i])=\begin{cases}c_C, & 1\le i\le 2\\ c_S, & 3\le i\le 38\\ c_N, & 39\le i\le 42\end{cases},$$

$$f_{CSBP}(M[i])=\begin{cases}c_C, & 1\le i\le 2\\ c_{Str}, & 3\le i\le 18\\ c_B, & 19\le i\le 38\\ c_P, & 39\le i\le 42\end{cases}.$$

对于簇分特征集合

$$\Delta\text{-}fea=\left\{\begin{array}{l}F\_56b,F\_128b,\\ F\_192b,F\_256b,\\ F\_56cut7,F\_128cut16,\\ F\_192cut24,F\_256cut32,\\ F\_256\end{array}\right\},$$

实验考察在提取不同簇分特征情况下（即 fea_{CM}，fea_{CSN}，$fea_{CSBP}\in\Delta\text{-}fea$），以准确率 p_{CM}^{O}、p_{CSN}^{O}、p_{CSBP}^{O} 为指标，簇分识别方案的表现. 对纳入考虑的数据，进行重复随机子抽样验证（Repeat random sub-sampling validation），每次抽样包含对每一密码体制密文文件的特征数据及相应标签进行抽样，随机抽样 80% 作为训练集，其余 20% 特征数据作为测试集. 以十折重复随机子抽样验证中测试集识别的平均准确

率作为簇分识别效果的度量. 3 种簇分设置 Δ_{CM}、Δ_{CSN}、Δ_{CSBP} 在不同簇分特征下的识别结果见表 3.

表 3 提取不同簇分特征时 3 种簇分设置的识别准确率

簇分特征 Δ-fea	准确率 p^O_{CM}/%	准确率 p^O_{CSN}/%	准确率 p^O_{CSBP}/%
随机分类	50.00	33.33	25.00
F_56b	100.00	93.88	64.64
F_56cut7	100.00	93.37	61.85
F_128b	100.00	98.72	69.36
$F\_128cut16$	100.00	96.72	66.97
F_192b	100.00	97.84	68.42
$F\_192cut24$	100.00	95.95	66.45
F_256b	**100.00**	**98.81**	**73.29**
$F\_256cut32$	100.00	97.87	68.36
F_256	100.00	93.93	64.65

实验表明，在 Δ_{CM}、Δ_{CSN}、Δ_{CSBP} 三种簇分设置下，当特征 $fea_{CM}, fea_{CSN}, fea_{CSBP} \in \Delta\text{-}fea$ 时，簇分识别的准确率均明显高于随机分类的准确率. 其中，最高准确率在 $fea_{CM}=fea_{CSN}=fea_{CSBP}=F\_256b$ 时达到，分别为 $p^O_{CM}=100.00\%$、$p^O_{CSN}=98.81\%$、$p^O_{CSBP}=73.29\%$.

由于 p^O_{CM} 与 p^O_{CSN} 的值接近 100.00%，根据实验结果可知，有理由认为 *CM*-簇分与 *CSN*-簇分具有可行性，亦即，密码体制类别划分(1)、(2)(第 3.2 节)确实在密码体制生成的密文数据中有所反映.

簇分设置 Δ_{CSBP} 下，最优准确率为 $p^O_{CSBP}=73.29\%$，虽然相比随机分类具有较大优势，但与前两种簇分设置下接近 100% 的准确率存在一定差距. 基于下述考虑，对其识别结果进行详细分析，以便进一步开展研究：(1) 相比 *CM*-簇分和 *CSN*-簇分，*CSBP*-簇分分类更加细致，更适于包含较多密码体制的分层识别任务(簇类别越多，相应每一簇内单分需要识别的体制就越少，从而降低识别的难度)；(2) *CSBP*-簇分识别结果相比随机分类仍具有较大优势，存在改进和发现的可能.

表 4 显示了 $fea_{CSBP}=F\_256b$ 时，簇分设置 Δ_{CSBP} 下的识别分类情况(因十折验证情形相仿，选取最后一折展示). 可见，古典簇 c_C、分组簇 c_B 中的测试样本都被识别到了其原本对应的簇中，而序列簇 c_{Str}、公钥簇 c_P 中有部分测试样本被识别到分组簇 c_B 中，尤其是序列簇 c_{Str} 中只有约 1/3 的样本被识别到其对应的簇中. 因此，在考察簇分结果的细节时，有必要对序列簇 c_{Str}、公钥簇 c_P 中的识别情况进行进一步分析.

表 4 特征 *F_256b* 在簇分设置 Δ_{CSBP} 下的识别结果

	古典簇 c_C	序列簇 c_{Str}	分组簇 c_B	公钥簇 c_P
古典簇 c_C	400	0	0	0
序列簇 c_{Str}	0	1092	2108	0
分组簇 c_B	0	0	4000	0
公钥簇 c_P	0	0	104	496

取出序列簇 c_{Str}、公钥簇 c_P 的识别结果，发现在序列簇 c_{Str} 中：(1) 属于 4 种 Trivium 体制的测试样本中有 24 个(共计 800 个)被正确识别，占这 4 种体制样本的 3.00%；(2) 属于 4 种 Grain 体制的测试样本中有 767 个(共计 800 个)被正确识别，占这 4 种体制样本的 95.88%；(3) 属于 4 种 RC4 体制的测试样本中有 301 个(共计 800 个)被正确识别，占这 4 种体制样本的 37.63%；(4) 所有属于 4 种 Sosemanuk 体制的测试样本(共计 800 个)都被识别到分组簇 c_B 中. 另一方面，在公钥簇 c_P 中：(1) 属于 2 种 RSA 体制的测试样本中有 388 个(共计 400 个)被正确识别，占这 2 种体制样本的 97.00%；(2) 属于 2 种 ECC 体制的测试样本中有 108 个(共计 200 个)被正确识别，占这 2 种体制样本的 54.00%.

考察上述结果，提出以下几种猜测：(1) 密码体制 T1、T2、T3、T4、So1、So2、So3、So4 与分组簇 c_B 对应的密码体制在所提取的密文特征上不存在明显差异；(2) 密码体制 rc4-1、rc4-2、rc4-3、rc4-4 与分组簇 c_B 对应的密码体制在所提取的密文特征上存在一定差异，但差异并不普遍；(3) 密码体制 G1、G2、G3、G4 与古典簇 c_C、分组簇 c_B、公钥簇 c_P 对应的密码体制在所提取的密文特征上存在较明显差异；(4) 密码体制 R1、R2 与古典簇 c_C、序列簇 c_{Str}、分组簇 c_B 对应的密码体制在所提取的密文特征上存在较明显差异；(5) 密码体制 E1、E2 与分组簇 c_B 对应的密码体制在所提取的密文特征上存在一定差异，但差异并不普遍.

结合簇分设置 Δ_{CSBP} 下的识别结果和上述猜测，既得到了一般意义下 *CSBP*-簇分可行的有利证据，也得到了一些相反的可能性. 因讨论对象有限，这一问题还有待更深入的研究. 接下来，为进一步开展对 *CSBP*-簇分的研究，主要对猜测(3)、(4)进行考察.

为此，取簇分设置

$$\Delta_{CSBP,1}=(M_1, C_{CSBP}, f_{CSBP,1}, O_{CSBP,1}, p^O_{CSBP,1}),$$

其中

$$M_1=\left\{\begin{array}{l}\{\mathrm{Sub},\mathrm{Perm}\}\\ \{\mathrm{G1},\mathrm{G2},\mathrm{G3},\mathrm{G4}\}\\ \{\mathrm{A1},\mathrm{A2},\mathrm{A3},\mathrm{A4},\mathrm{A5},\mathrm{A6}\}\\ \{\mathrm{C1},\mathrm{C2},\mathrm{C3},\mathrm{C4},\mathrm{C5},\mathrm{C6}\}\\ \{\mathrm{D1\text{-}1},\mathrm{D1\text{-}2},\mathrm{D3\text{-}1},\mathrm{D3\text{-}2}\}\\ \{\mathrm{S1},\mathrm{S2},\mathrm{S3},\mathrm{S4}\}\\ \{\mathrm{R1},\mathrm{R2}\}\end{array}\right\},$$

$f_{CSBP,1}=f_{CSBP}\mid_{M_1}$，$O_{CSBP,1}=(CPO,F_256b,RF)$进行实验，即在$\Delta_{CSBP}$其余设置不变的前提下，在序列簇$c_{Str}$中仅保留密码体制 G1、G2、G3、G4，在公钥簇c_P中仅保留密码体制 R1、R2，开展*CSBP*-簇分实验.

实验结果显示，在簇分设置$\Delta_{CSBP,1}$下，准确率$p^O_{CSBP,1}=99.99\%$，其分类情况见表 5（选取十折验证中最后一折展示）.

表 5　*CSBP*-簇分设置$\Delta_{CSBP,1}$下识别的分类情况

	古典簇c_C	序列簇c_{Str}	分组簇c_B	公钥簇c_P
古典簇c_C	400	0	0	0
序列簇c_{Str}	0	792	8	0
分组簇c_B	0	0	4000	0
公钥簇c_P	0	0	0	400

由准确率及表 5 知，在序列簇c_{Str}仅保留 4 种 Grain 密码体制、在公钥簇c_P中仅保留 2 种 RSA 密码体制的情形下，*CSBP*-簇分被较好地完成. 这意味着，虽然目前无法在一般意义下断言 *CSBP*-簇分的可行性，但发现了其可行的一种情形.

5.3　密码体制的分层识别

基于第 5.2 节发现的 *CSBP*-簇分可行情形，本节讨论密码体制的分层识别.

5.3.1　簇分设置及识别结果

簇分设置$\Delta_{CSBP,1}$中，在分组簇c_B原有的 20 种密码体制被完全保留的情况下，仍取得了较好的簇分识别效果. 这意味着在分层识别中，对分组簇c_B中的密码体制进行一定的组织筛选，不会对簇分识别效果产生大的影响.

为降低分组簇c_B中密码体制的无序性，同时在分层识别中，揭示分组簇c_B内单分中可能存在的规律和问题，按照分组簇c_B内考虑：(1) 相同工作模式的不同分组密码体制；(2) 不同工作模式、相同分组长度的同一分组密码体制；(3) 相同工作模式、不同分组长度的同一分组密码体制的 3 个原则，针对性地选择了如下 3 种设置开展分层识别研究：

(1) *CSBP*-簇分设置

$\Delta_{CSBP,2}=(M_2,C_{CSBP},f_{CSBP,2},O_{CSBP,2},p^O_{CSBP,2})$，其中

$$M_2=\begin{Bmatrix}\text{Sub,Perm}\\ \text{G1,G2,G3,G4}\\ \text{A1,C1,D1-1,D3-1,S1}\\ \text{R1,R2}\end{Bmatrix},$$

$f_{CSBP,2}=f_{CSBP}\mid_{M_2}$，$O_{CSBP,2}=(CPO,F\_256b,RF)$.

(2) *CSBP*-簇分设置

$\Delta_{CSBP,3}=(M_3,C_{CSBP},f_{CSBP,3},O_{CSBP,3},p^O_{CSBP,3})$，其中

$$M_3=\begin{Bmatrix}\text{Sub,Perm}\\ \text{G1,G2,G3,G4}\\ \text{S1,S2,S3,S4}\\ \text{R1,R2}\end{Bmatrix},$$

$f_{CSBP,3}=f_{CSBP}\mid_{M_3}$，$O_{CSBP,3}=(CPO,F\_256b,RF)$.

(3) *CSBP*-簇分设置

$\Delta_{CSBP,4}=(M_4,C_{CSBP},f_{CSBP,4},O_{CSBP,4},p^O_{CSBP,4})$，其中

$$M_4=\begin{Bmatrix}\text{Sub,Perm}\\ \text{G1,G2,G3,G4}\\ \text{A1,A3,A5}\\ \text{R1,R2}\end{Bmatrix},$$

$f_{CSBP,2}=f_{CSBP}\mid_{M_2}$，$O_{CSBP,2}=(CPO,F\_256b,RF)$.

实验显示，上述 *CSBP*-簇分设置下，准确率为$p^O_{CSBP,2}=99.70\%$、$p^O_{CSBP,3}=99.78\%$、$p^O_{CSBP,4}=99.79\%$.

由定义 1、2，现在讨论密码体制识别问题

$$\Gamma_{M_2}=(M_2,J_{M_2},h^{J_{M_2}})、$$

$$\Gamma_{M_3}=(M_3,J_{M_3},h^{J_{M_3}})、$$

$$\Gamma_{M_4}=(M_4,J_{M_4},h^{J_{M_4}})$$

下，分层识别方案

$$J_{M_2}=(TLSO,FEA_{M_2},RF)、$$

$$J_{M_3}=(TLSO,FEA_{M_3},RF)、$$

$$J_{M_4}=(TLSO,FEA_{M_4},RF)$$

的识别效果.

由于Γ_{M_2}、Γ_{M_3}、Γ_{M_4}对应分层识别方案中方案工作流程 *TLSO* 的识别情境下的簇分设置$\Delta_{CSBP,2}$、$\Delta_{CSBP,3}$、$\Delta_{CSBP,4}$及在这 3 种簇分设置下的识别准确率$p^O_{CSBP,2}$、$p^O_{CSBP,3}$、$p^O_{CSBP,4}$已知，要探讨分层识别方案的整体识别效果，在此基础之上考虑相应单分设置的识别效果，再对 2 阶段识别结果进行整合即可.

5.3.2　单分设置及结果分析

记Γ_{M_2}、Γ_{M_3}、Γ_{M_4}对应分层识别方案中 *TLSO* 的识别情境为

$$S_{M_2}=(\Delta_{CSBP,2},4,\Theta_1^{M_2},\Theta_2^{M_2},\Theta_3^{M_2},\Theta_4^{M_2})、$$

$$S_{M_3}=(\Delta_{CSBP,3},4,\Theta_1^{M_3},\Theta_2^{M_3},\Theta_3^{M_3},\Theta_4^{M_3})、$$

$$S_{M_4}=(\Delta_{CSBP,4},4,\Theta_1^{M_4},\Theta_2^{M_4},\Theta_3^{M_4},\Theta_4^{M_4}),$$

则对其中的单分设置，有

$$\Theta_1^{M_2}=\Theta_1^{M_3}=\Theta_1^{M_4}=(\Delta_{CSBP,1},\{\text{Sub,Perm}\},c_C,U_1,q^{U_1}) \tag{1}$$

$$\Theta_2^{M_2}=\Theta_2^{M_3}=\Theta_2^{M_4}=(\Delta_{CSBP,1},\{\text{G1,G2,G3,G4}\},c_{Str},U_2,q^{U_2}) \tag{2}$$

$$\Theta_4^{M_2}=\Theta_4^{M_3}=\Theta_4^{M_4}=(\Delta_{CSBP,1},\{\mathrm{R1},\mathrm{R2}\},c_P,U_4,q^{U_4}) \quad (3)$$

$$\Theta_3^{M_2}=(\Delta_{CSBP,2},\{\mathrm{A1},\mathrm{C1},\mathrm{D1\text{-}1},\mathrm{D3\text{-}1},\mathrm{S1}\},c_B,U_3^{M_2},q^{U_3^{M_2}}) \quad (4)$$

$$\Theta_3^{M_3}=(\Delta_{CSBP,3},\{\mathrm{S1},\mathrm{S2},\mathrm{S3},\mathrm{S4}\},c_B,U_3^{M_3},q^{U_3^{M_3}}) \quad (5)$$

$$\Theta_3^{M_4}=(\Delta_{CSBP,4},\{\mathrm{A1},\mathrm{A3},\mathrm{A5}\},c_B,U_3^{M_4},q^{U_3^{M_4}}) \quad (6)$$

式(1)～(3)中,由于 $\Delta_{CSBP,1}$、$\Delta_{CSBP,2}$、$\Delta_{CSBP,3}$、$\Delta_{CSBP,4}$下古典簇 c_C、序列簇 c_{Str}、公钥簇 c_P下的单分具有相同的含义,因此直接用 $\Delta_{CSBP,1}$表示,将它们等同起来.式(4)～(6)中,$U_3^{M_2}$、$U_3^{M_3}$、$U_3^{M_4}$的不同意味着对分组簇 c_B对应的不同密码体制集合,采取不同的单分特征进行单分识别是可能的.

在式(1)～(6)所述单分设置下,对可能提取的单分特征集合

$$\Theta\text{-}fea=\left\{\begin{array}{l}Max,\\ F\_F,F\_BF,F\_R,F\_3Test,\\ Ent,Max+Ent,\\ F\_512,F\_1024,\\ F\_5Test,F\_256b,F\_256\end{array}\right\}$$

进行了实验,得各单分设置下的识别效果(见表 6).

表 6 各单分设置下不同特征的识别准确率

单分特征	准确率 q^{U_1}/%	准确率 q^{U_2}/%	准确率 q^{U_4}/%	准确率 $q^{U_3^{M_2}}$/%	准确率 $q^{U_3^{M_3}}$/%	准确率 $q^{U_3^{M_4}}$/%
F_256b	**100.00**	35.88	**98.85**	**21.15**	24.88	32.76
F_256	—	35.63	—	20.52	24.46	32.90
Max	—	25.00	—	20.00	23.98	33.33
F_F	—	17.38	—	17.10	21.88	29.00
F_BF	—	17.88	—	16.00	21.00	31.17
F_R	—	16.68	—	18.10	20.63	27.67
F_3Test	—	17.13	—	17.90	22.38	30.50
Ent	—	34.88	—	19.79	**26.63**	34.29
Max+Ent	—	35.50	—	19.01	24.55	34.68
F_5Test	—	36.00	—	20.75	25.10	34.33
F_512	—	**36.25**	—	19.58	25.00	34.57
F_1024	—	35.63	—	19.40	26.05	**34.73**
随机分类	50.00	25.00	50.00	20.00	25.00	33.33

首先沿用簇分识别中表现最好的特征 *F_256b* 进行了各单分设置下的识别.在式(1)、式(3)给出的古典簇单分及公钥簇单分下,识别准确率分别达到 100.00%和 99.85%,认为其基本达到进行准确识别的要求.而在由式(2)、式(4)～(6)给出的序列簇单分和分组簇单分下,识别准确率仍存在改善空间.因此,使用 Θ-*fea* 中的其它特征进行单分识别时,仅在由式(2)、式(4)～(6)给出的序列簇单分和分组簇单分下进行实验,并列出相关结果.

由表 6 知,在单分设置 $\Theta_2^{M_2}=\Theta_2^{M_3}=\Theta_2^{M_4}$(序列簇的单分)下,特征 $fea^{U_2}=F\_512$ 时,达到最高准确率 $q^{U_2}=36.25\%$;在单分设置 $\Theta_3^{M_2}$(簇分设置 $\Delta_{CSBP,2}$下的分组簇单分)下,特征 $fea^{U_3^{M_2}}=F\_256b$ 时,达到最高准确率 $q^{U_3^{M_2}}=21.15\%$;在单分设置 $\Theta_3^{M_3}$(簇分设置 $\Delta_{CSBP,3}$下的分组簇单分)下,特征 $fea^{U_3^{M_3}}=Ent$ 时,达到最高准确率 $q^{U_3^{M_3}}=26.63\%$;在单分设置 $\Theta_3^{M_4}$(簇分设置 $\Delta_{CSBP,4}$下的分组簇单分)下,特征 $fea^{U_3^{M_4}}=F\_1024$ 时,达到最高准确率 $q^{U_3^{M_4}}=34.73\%$.

由于上述结果中,3 个分组簇内单分的准确率 $q^{U_3^{M_2}}=21.15\%$、$q^{U_3^{M_3}}=26.63\%$、$q^{U_3^{M_4}}=34.73\%$ 与随机分类的准确率相差较小,在对簇分阶段与单分阶段识别结果进行整合、分析分层识别方案的整体识别效果前,提取相应实验中所有十折验证过程的准确率结果,通过单样本 *T*-检验检验其准确率是否大于随机准确率.经过检验,得到 p 值分别为 0.0117、0.0014、0.02424,均小于显著性水平 0.05.假设检验结果表明,在单分设置 $\Theta_3^{M_2}$、$\Theta_3^{M_3}$、$\Theta_3^{M_4}$下,使用特征 $fea^{U_3^{M_2}}=F\_256b$、$fea^{U_3^{M_3}}=Ent$、$fea^{U_3^{M_4}}=F\_1024$ 进行识别时,其准确率在统计学意义下显著大于随机分类准确率.因此,认为选择的几种特征是有意义的.

5.3.3 分层识别总体评析

现在整合簇分与单分阶段识别结果,对密码体制识别问题 Γ_{M_2}、Γ_{M_3}、Γ_{M_4}下分层识别方案

$$J_{M_2}=(TLSO,FEA_{M_2},RF)、$$

$$J_{M_3}=(TLSO,FEA_{M_3},RF)、$$

$$J_{M_4}=(TLSO,FEA_{M_4},RF)$$

的识别效果进行评价.

所讨论的 3 个分层识别方案的识别准确率由式 $h^{J_{M_2}}=p^O_{CSBP,2}\cdot(r_C^2\cdot q^{U_1}+r_{Str}^2\cdot q^{U_2}+r_B^2\cdot q^{U_3^{M_2}}+r_P^2\cdot q^{U_4})$、$h^{J_{M_3}}=p^O_{CSBP,3}\cdot(r_C^3\cdot q^{U_1}+r_{Str}^3\cdot q^{U_2}+r_B^3\cdot q^{U_3^{M_3}}+r_P^3\cdot q^{U_4})$、

$h^{J_{M_4}}=p^{O}_{CSBP,4}\cdot(r^4_C\cdot q^{U_1}+r^4_{Str}\cdot q^{U_2}+r^4_B\cdot q^{U_3^{M_4}}+r^4_P\cdot q^{U_4})$ 给出，其中 r^i_C、r^i_{Str}、r^i_B、r^i_P 分别代表密码体制识别问题 Γ_{M_i} 的簇分设置 $\Delta_{CSBP,i}$ 下，c_C、c_{Str}、c_B、c_P 各簇中测试样本占总测试样本的比例，有

$r^2_C=2/13$，$r^2_{Str}=4/13$，$r^2_B=5/13$，$r^2_P=2/13$，

$r^3_C=2/12$，$r^3_{Str}=4/12$，$r^3_B=4/12$，$r^3_P=2/12$，

$r^4_C=2/11$，$r^4_{Str}=4/11$，$r^4_B=3/11$，$r^4_P=2/11$.

表 7 列出了按上述式子计算得到的 3 个分层识别方案的识别准确率，分别为 $h^{J_{M_2}}=49.88\%$、$h^{J_{M_3}}=54.15\%$、$h^{J_{M_4}}=58.87\%$.

表 7　双层识别方案与单层识别方案的准确率比较

来源文献	使用特征	方案结构	准确率 $h^{J_{M_2}}$/%	准确率 $h^{J_{M_3}}$/%	准确率 $h^{J_{M_4}}$/%
—	随机分类	单层	7.69	8.33	9.09
[5]	*Max*	单层	7.69	8.33	9.09
本文	*Ent*	单层	24.43	26.54	29.22
本文,[5]	*Max*+*Ent*	单层	25.18	26.60	29.64
[3,6]	*F_256*	单层	30.18	32.58	35.69
本文	*F_256b*,*F_512*	分层	**49.73**	—	—
本文	*F_256b*,*F_512*,*Ent*	分层	—	**53.98**	—
本文	*F_256b*,*F_512*,*F_1024*	分层	—	—	**58.68**

同时，进行实验，与选用不同特征的单层识别方案进行了比较. 据表 7 可知：(1) 使用不同特征集的单层识别方案，在识别准确率上存在着一定差异；(2) 选用特征 *Max* 的单层识别方案不具备优于随机分类的密码体制识别能力(进一步观察所提取特征，了解到这是由于对于长度达到 512 KB 的不同密文文件，提取到的特征 *Max* 的值相同而导致的)；(3) 单层识别方案中，使用特征 *Ent*、*Max*+*Ent*、*F_256* 的 3 种单层方案效果识别准确率明显高于随机分类；(4) 鉴于(2)，认为表 7 中选用特征 *Ent* 与 *Max*+*Ent* 的单层识别方案在准确率上不存在显著差异；(5) 单层识别方案中，选用特征 *F_256* 的方案在 3 个密码体制识别问题下都取得最高的准确率，分别为 $h^{J_{M_2}}=30.18\%$、$h^{J_{M_3}}=32.58\%$、$h^{J_{M_4}}=35.69\%$；(6) 对密码体制识别问题 Γ_{M_2}、Γ_{M_3}、Γ_{M_4}，与单层识别方案相比，分层识别方案均具有较明显的优势(其准确率分别较单层识别方案的最优结果提升 19.55%、21.40%、22.99%).

由上述分析，在密码体制识别中，所提分层识别方案优于单层识别方案. 为进一步说明，对 3 种情况下分层识别方案准确率(十折)与最优单层识别方案的准确率(十折)进行独立两样本 T-检验. 检验结果显示，3 个 p 值均小于 2.2×10^{-16}，表明所得结论在统计学意义下是显著的.

6　总结与展望

本文通过对密码体制识别问题的探讨，给出了一个密码体制识别问题的完整定义系统，使得单层识别方案和所提基于随机森林的密码体制分层识别方案都能用简洁统一的方式描述. 考虑到密码体制识别问题相对于一般模式识别问题的特殊性，该定义系统具有足够的包容性，能够为未来密码体制识别研究提供一个一致的、便于整合的描述框架，具有积极意义. 同时，由于密码体制识别问题的定义系统尚属首次提出，可能仍存在值得改进、完善之处，有待未来研究探讨.

为开展密码体制的分层识别，定义了 3 种具体簇分，并探讨了其可行性. 分析结果认为，*CM*-簇分与 *CSN*-簇分具有较强可行性，而一般意义下 *CSBP*-簇分的可行性有待进一步研究验证. 由于 *CSBP*-簇分在 3 种具体簇分中具有更明显的现实意义，对于这一问题的深入研究是值得开展的.

进一步，发现了 *CSBP*-簇分可行的一种情形，在该情形下，开展了密码体制的分层识别. 对应 3 种不同设置，运用提出的基于随机森林的密码体制分层识别方案开展了识别实验. 实验结果显示，与单层识别方案相比，所提分层识别方案在对识别准确率的作用上有显著的提升. 因此，未来关于密码体制识别的研究可参考本文提出的分层识别模式.

虽然分层识别方案的整体识别效果较好，但单分阶段的实验细节显示，在分组簇和序列簇内进行单分的识别效果还存在较大改进空间. 特别地，发现：(1) 采取相同工作模式的不同分组密码体制较难识别；(2) 采取不同工作模式的相同分组密码体制较难识别；(3) 选取不同分组长度的相同分组密码体制较难识别. 针对上述问题，未来研究可能的思路有：(1) 对相同工作模式的不同密码体制的体制结构进行深入研究，提取与之相关的特征来进行识别；(2) 对采取不同工作模式的相同密码体制，研究

不同工作模式的处理流程,提取与之相关的特征来进行识别;(3)对不同分组长度的相同密码体制,考察其对不同分组长度设置进行加密操作的差异,提取与之相关的特征来进行识别.另外,在公钥簇内的单分取得了较好的识别结果,然而所考虑公钥密码体制较少,仍有待后续研究将结果向一般情形推广.

致 谢 感谢审稿专家提出的宝贵意见,这些意见对论文的完善提供了很大帮助.同时感谢信息工程大学魏江宏博士在论文修改期间的讨论和建议!

参 考 文 献

[1] Wu Yang, Wang Tao, Xing Meng, et al. Blockciphers identification scheme based on the distribution character of randomness test values of ciphertext. Chinese Journal on Communications, 2015, 36(4): 146-155(in Chinese)
(吴杨,王韬,邢萌等. 基于密文随机性度量值分布特征的分组密码算法识别方案. 通信学报, 2015, 36(4): 146-155)

[2] De Souza W A R, Tomlinson A. A distinguishing attack with a neural network//Proceedings of the IEEE 13th International Conference on Data Mining Workshops(ICDMW'13). Dallas, USA, 2013: 154-161

[3] Sharif S O, Mansoor S P. Performance evaluation of classifiers used for identification of encryption algorithms. ACEEE International Journal on Network Security, 2011, 2(4): 42-45

[4] Dileep A D, Sekhar C C. Identification of block ciphers using support vector machines//Proceedings of the International Joint Conference on Neural Networks(IJCNN'06). Gulf Islands, Canada, 2006: 2696-2701

[5] Manjula R, Anitha R. Identification of encryption algorithm using decision tree//Proceedings of the 1st International Conference on Computer Science and Information Technology. Bangalore, India, 2011: 237-246

[6] Sharif S O, Kuncheva L I, Mansoor S P. Classifying encryption algorithms using pattern recognition techniques//Proceedings of the IEEE International Conference on Information Theory and Information Security(ICITIS). Beijing, China, 2010: 1168-1172

[7] Nagireddy S. A Pattern Recognition Approach to Block Cipher Identification[M. S. dissertation]. Indian Institute of Technology Madras, India, 2008

[8] Lomte V M, Shinde A D. Review of a new distinguishing attack using block cipher with a neural network. International Journal of Science and Research, 2014, 3(8): 733-736

[9] de Souza W A R, de Carvalho L A V, Xexéo J. Identification of N Block Ciphers. IEEE Latin America Transactions, 2011, 9(2): 184-191

[10] Chou J W, Lin S D, Cheng C M. On the effectiveness of using state-of-the-art machine learning techniques to launch cryptographic distinguishing attacks//Proceedings of the 5th ACM Workshop on Security and Artificial Intelligence. New York, USA, 2012: 105-110

[11] Turan M S, Çahk Ç, Saran N B, et al. New distinguishers based on random mappings against stream ciphers//Proceedings of the 15th International Conference on Sequences and Their Applications(SETA'08). Lexington, USA, 2008: 30-41

[12] Mishra S, Bhattacharjya A. Pattern analysis of cipher text: A combined approach//Proceedings of the 3rd International Conference on Recent Trends in Information Technology(ICRTIT'13). Chennai, India, 2013: 393-398

[13] Liaw A, Wiener M. Classification and regression by randomForest. R News, 2002, 2(3): 18-22

[14] Douglas R Stinson, Feng Deng-Guo. Cryptography Theory and Practice. 3rd Edition. Beijing: Publishing House of Electronics Industry, 2016(in Chinese)
(斯廷森,冯登国. 密码学原理与实践. 第 3 版. 北京:电子工业出版社, 2016)

[15] Chen Shao-Zhen. Cryptography Tutorial. Beijing: Science Press, 2012(in Chinese)
(陈少真. 密码学教程. 北京:科学出版社, 2012)

[16] Yang Yi-Xian, Niu Xin-Xin. Applied Cryptography. Beijing: Beijing University of Posts and Telecommunications Press, 2013(in Chinese)
(杨义先,钮心忻. 密码学教程. 北京:北京邮电大学出版社, 2013)

[17] Wu Wen-Ling, Feng Deng-Guo, Zhang Wen-Tao. The Design and Analysis of Block Cipher. Beijing: Tsinghua University Press, 2009(in Chinese)
(吴文玲,冯登国,张文涛. 分组密码的设计与分析. 北京:清华大学出版社, 2009)

[18] Rukhin A, Soto J, Nechvatal J, et al. A statistical test suite for random and pseudorandom number generators for cryptographic applications. Gaithersburg, USA: Booz-Allen and Hamilton Inc Mclean Va, NIST Special Publication, 2001

附 录.

簇分阶段和单分阶段涉及到 19 种特征的提取,在此对其中计算细节进行介绍.

首先给出一些将要用到的标记和定义.

对于一个待提取特征的文件(记为 $File$)来说,可将其简单地视为一个比特串 $\boldsymbol{B}=(b_1,b_2,\cdots,b_{l_B})$或一个字节串 $\boldsymbol{C}=(c_1,c_2,\cdots,c_{l_C})$,其中 l_B 为其比特串长度,l_C 为其字节串长度,满足 $l_B=8\cdot l_C$.

熵在信息论中用来度量信息的不确定性和数据的随机

性.对随机变量 X,假设其有 k 种可能的取值 $x_1,\cdots,x_k$,第 i 种取值 x_i 出现的概率为 p_i.则随机变量 X 的熵定义如下:

$$H(X)=-\sum_{i=1}^{k}p_i\log p_i.$$

若随机变量 X 所有 k 个可能取值 $x_1,\cdots,x_k$ 出现概率相等,即 $p_1=p_2=\cdots=p_k=p=1/k$,则其值熵达到最大,称为最大熵,计算公式如下:

$$H_{Max}(X)=-k\cdot p\cdot\log p=\log k.$$

以下所有概率计算均以频率近似.

(1)特征 F_56b

56 维特征 F_56b 形如 $(f_1^{56},f_2^{56},\cdots,f_{56}^{56})$,计算为

$$f_i^{56}=-p_i^{56}\log p_i^{56}-(1-p_i^{56})\log(1-p_i^{56}),\ i=1,\cdots,56,$$

其中 $p_i^{56}=\dfrac{\sum_{j=1}^{\left\lfloor\frac{l_B}{56}\right\rfloor}b_{i+56\cdot(j-1)}}{\left\lfloor\frac{l_B}{56}\right\rfloor},\ i=1,\cdots,56.$

(2)特征 F_128b

128 维特征 F_128b 形如 $(f_1^{128},f_2^{128},\cdots,f_{128}^{128})$,计算为

$$f_i^{128}=-p_i^{128}\log p_i^{128}-(1-p_i^{128})\log(1-p_i^{128}),\quad i=1,\cdots,128,$$

其中 $p_i^{128}=\dfrac{\sum_{j=1}^{\left\lfloor\frac{l_B}{128}\right\rfloor}b_{i+128\cdot(j-1)}}{\left\lfloor\frac{l_B}{128}\right\rfloor},\ i=1,\cdots,128.$

(3)特征 F_192b

192 维特征 F_192 形如 $(f_1^{192},f_2^{192},\cdots,f_{192}^{192})$,计算为

$$f_i^{192}=-p_i^{192}\log p_i^{192}-(1-p_i^{192})\log(1-p_i^{192}),\quad i=1,\cdots,192,$$

其中 $p_i^{192}=\dfrac{\sum_{j=1}^{\left\lfloor\frac{l_B}{192}\right\rfloor}b_{i+192\cdot(j-1)}}{\left\lfloor\frac{l_B}{192}\right\rfloor},\ i=1,\cdots,192.$

(4)特征 F_256b

256 维特征 F_256 形如 $(f_1^{256},f_2^{256},\cdots,f_{256}^{256})$,计算为

$$f_i^{256}=-p_i^{256}\log p_i^{256}-(1-p_i^{256})\log(1-p_i^{256}),\quad i=1,\cdots,256,$$

其中 $p_i^{256}=\dfrac{\sum_{j=1}^{\left\lfloor\frac{l_B}{256}\right\rfloor}b_{i+256\cdot(j-1)}}{\left\lfloor\frac{l_B}{256}\right\rfloor},\ i=1,\cdots,256.$

(5)特征 F_56cut7

7 维特征 F_56cut7 形如 $(f_1^{56cut7},f_2^{56cut7},\cdots,f_7^{56cut7})$,计算为

$$f_i^{56cut7}=-\sum_{j=1}^{256}p_{i,j}^{7X256}\log p_{i,j}^{7X256},\ i=1,\cdots,7,$$

其中 $p_{i,j}^{7X256}$ 表示分块中第 i 个固定字节位取值为 $j-1$ 的字节出现的概率.

(6)特征 $F\_128cut16$

16 维特征 $F\_128cut16$ 形如 $(f_1^{128cut16},f_2^{128cut16},\cdots,f_{16}^{128cut16})$,计算为

$$f_i^{128cut16}=-\sum_{j=1}^{256}p_{i,j}^{16X256}\log p_{i,j}^{16X256},\ i=1,\cdots,16,$$

其中 $p_{i,j}^{16X256}$ 表示分块中第 i 个固定字节位取值为 $j-1$ 的字节出现的概率.

(7)特征 $F\_192cut24$

24 维特征 $F\_192cut24$ 形如 $(f_1^{192cut24},f_2^{192cut24},\cdots,f_7^{192cut24})$,计算为

$$f_i^{192cut24}=-\sum_{j=1}^{256}p_{i,j}^{24X256}\log p_{i,j}^{24X256},\ i=1,\cdots,24,$$

其中 $p_{i,j}^{24X256}$ 表示分块中第 i 个固定字节位取值为 $j-1$ 的字节出现的概率.

(8)特征 $F\_256cut32$

32 维特征 $F\_256cut32$ 形如 $(f_1^{256cut32},f_2^{256cut32},\cdots,f_{32}^{256cut32})$,计算为

$$f_i^{256cut32}=-\sum_{j=1}^{256}p_{i,j}^{32X256}\log p_{i,j}^{32X256},\ i=1,\cdots,32,$$

其中 $p_{i,j}^{32X256}$ 表示分块中第 i 个固定字节位取值为 $j-1$ 的字节出现的概率.

(9)特征 F_256

256 维特征 F_256 形如 $(f_1^{256},\cdots,f_{256}^{256})$,计算为

$$f_i^{256}=\frac{N_i^{256}}{l_C},\ i=1,\cdots,256,$$

其中,N_i^{256} 表示在 $File$ 中,取值为 $i-1$ 的字节出现的频数.

(10)特征 Max

6 维特征 Max 形如 $(f_1^M,f_2^M,\cdots,f_6^M)$,计算为

$$f_1^M=\log n_1^M,\ f_2^M=\log n_2^M,\ f_3^M=\log n_3^M,$$
$$f_4^M=\log n_4^M,\ f_5^M=\log n_5^M,\ f_6^M=\log n_6^M,$$

其中 n_1^M 为 $File$ 中 256 种字节取值出现的个数(对任一字节取值,只要在 $File$ 中出现一次,则被计入 n_1^M 中,以下类似);n_2^M 为 $File$ 中 31～255 共 224 种字节取值出现的个数;n_3^M 为 $File$ 中 65～90 共 26 种字节取值出现的个数;n_4^M 为 $File$ 中 97～122 共 26 种字节取值出现的个数;n_5^M 为 $File$ 中 48～57 共10 种字节取值出现的个数;n_6^M 为 $File$ 中 S_{194} 共 194 种字节取值出现的个数;

(11)特征 F_F

500 维特征 F_F 形如 $(f_1^F,\cdots,f_{500}^F)$,在特征提取时,将 $File$ 分为长度相等的 10000 个分块,对每一分块运行随机性测试 Frequency Test[18],得到一列返回值 $\boldsymbol{P}^F=(p_1^F,p_2^F,\cdots,p_{10000}^F)$.取 $\Delta p^F=0.002$,$d_F=1/\Delta p^F$,计算:

① $\boldsymbol{VP}_1^F=(vp_{1,1}^F,vp_{1,2}^F,\cdots,vp_{1,d_F}^F)$,其中 $vp_{1,i}^F=n_{1,i}^F\cdot\Delta p^F$,$n_{1,i}^F$ 为 $\boldsymbol{P}^{l,F}=(p_1^F,p_2^F,\cdots,p_{5000}^F)$ 中取值在 $[\Delta p^F\cdot(i-1),\Delta p^F\cdot i)$ 内的个数;

② $\boldsymbol{VP}_2^F=(vp_{2,1}^F,vp_{2,1}^F,\cdots,vp_{2,d_F}^F)$,其中 $vp_{2,i}^F=n_{2,i}^F\cdot\Delta p^F$,$n_{1,i}^F$ 为 $\boldsymbol{P}^{r,F}=(p_{5001}^F,p_{5002}^F,\cdots,p_{10000}^F)$ 中取值在 $[\Delta p^F\cdot(i-1),\Delta p^F\cdot i)$ 内的个数;

③$(f_1^F,\cdots,f_{500}^F)=|\boldsymbol{VP}_1^F-\boldsymbol{VP}_2^F|$.

(12)特征 *F_BF*

1000 维特征 *F_BF* 形如$(f_1^{BF},\cdots,f_{1000}^{BF})$,在特征提取时,将 *File* 分为长度相等的 10000 个分块,设置参数 Block Size M 为 20bits,对每一分块运行随机性测试 Block Frequency Test[18],得到一列返回值 $\boldsymbol{P}^{BF}=(p_1^{BF},p_2^{BF},\cdots,p_{10000}^{BF})$.取 $\Delta p^{BF}=0.001$,$d_{BF}=1/\Delta p^{BF}$,计算:

① $\boldsymbol{VP}_1^{BF}=(vp_{1,1}^{BF},vp_{1,2}^{BF},\cdots,vp_{1,d_F}^{BF})$,其中 $vp_{1,i}^{BF}=n_{1,i}^{BF}\cdot\Delta p^{BF}$,$n_{1,i}^{BF}$ 为 $\boldsymbol{P}^{l,BF}=(p_1^{BF},p_2^{BF},\cdots,p_{5000}^{BF})$ 中取值在$[\Delta p^{BF}\cdot(i-1),\Delta p^{BF}\cdot i)$内的个数;

② $\boldsymbol{VP}_2^{BF}=(vp_{2,1}^{BF},vp_{2,1}^{BF},\cdots,vp_{2,d_F}^{BF})$,其中 $vp_{2,i}^{BF}=n_{2,i}^{BF}\cdot\Delta p^{BF}$,$n_{1,i}^{BF}$ 为 $\boldsymbol{P}^{r,BF}=(p_{5001}^{BF},p_{5002}^{BF},\cdots,p_{10000}^{BF})$ 中取值在$[\Delta p^{BF}\cdot(i-1),\Delta p^{BF}\cdot i)$内的个数;

③$(f_1^{BF},\cdots,f_{500}^{BF})=|\boldsymbol{VP}_1^{BF}-\boldsymbol{VP}_2^{BF}|$.

(13)特征 *F_R*

2000 维特征 *F_R* 形如$(f_1^R,\cdots,f_{2000}^R)$,在特征提取时,将 *File* 分为长度相等的 10000 个分块,对每一分块运行随机性测试 Runs Test[18],得到一列返回值 $\boldsymbol{P}^{F\_R}=(p_1^R,p_2^R,\cdots,p_{10000}^R)$.取 $\Delta p^R=0.0005$,$d_R=1/\Delta p^R$,计算:

① $\boldsymbol{VP}_1^R=(vp_{1,1}^R,vp_{1,2}^R,\cdots,vp_{1,d_R}^R)$,其中 $vp_{1,i}^R=n_{1,i}^R\cdot\Delta p^R$,$n_{1,i}^R$ 为 $\boldsymbol{P}^{l,R}=(p_1^R,p_2^R,\cdots,p_{5000}^R)$ 中取值在$[\Delta p^R\cdot(i-1),\Delta p^R\cdot i)$内的个数;

② $\boldsymbol{VP}_2^R=(vp_{2,1}^R,vp_{2,2}^R,\cdots,vp_{2,d_R}^R)$,其中 $vp_{2,i}^R=n_{2,i}^R\cdot\Delta p^R$,$n_{1,i}^R$ 为 $\boldsymbol{P}^{r,R}=(p_{5001}^R,p_{5002}^R,\cdots,p_{10000}^R)$ 中取值在$[\Delta p^R\cdot(i-1),\Delta p^R\cdot i)$内的个数;

③$(f_1^R,\cdots,f_{500}^R)=|\boldsymbol{VP}_1^R-\boldsymbol{VP}_2^R|$.

(14)特征 *F_3Test*

特征 *F_3Test* 维数为 3500,是特征 *F_F*、*F_BF*、*F_R* 的组合,形如$(f_1^F,\cdots,f_{d_F}^F,f_1^{BF},\cdots,f_{d_{BF}}^{BF},f_1^R,\cdots,f_{d_R}^R)$.

(15)特征 *Ent*

6 维特征 *F_Ent* 形如$(f_1^E,\cdots,f_6^E)$,计算为

$$f_1^E=-\sum_{i=1}^{256}p_i^1\log p_i^1,\ f_2^E=-\sum_{i=1}^{224}p_i^2\log p_i^2,$$

$$f_3^E=-\sum_{i=1}^{26}p_i^3\log p_i^3,\ f_4^E=-\sum_{i=1}^{26}p_i^4\log p_i^4,$$

$$f_5^E=-\sum_{i=1}^{10}p_i^5\log p_i^5,\ f_6^E=-\sum_{i=1}^{194}p_i^6\log p_i^6,$$

其中 $p_i^1=\dfrac{N_i^1}{l_C}$,N_i^1 为取值为 $i-1$ 的字节在 *File* 中出现的频数,$i=1,\cdots,256$;$p_i^2=N_i^2\Big/\sum_{i=1}^{224}N_i^2$,$N_i^2$ 为取值为 $i+31$ 的字节在 *File* 中出现的频数,$i=1,\cdots,224$;$p_i^3=N_i^3\Big/\sum_{i=1}^{26}N_i^3$,$N_i^3$ 为取值为 $i+64$ 的字节在 *File* 中出现的频数,$i=1,\cdots,26$;$p_i^4=N_i^4\Big/\sum_{i=1}^{26}N_i^4$,$N_i^4$ 为取值为 $i+96$ 的字节在 *File* 中出现的频数,$i=1,\cdots,26$;$p_i^5=N_i^5\Big/\sum_{i=1}^{10}N_i^5$,$N_i^5$ 为取值为 $i+47$ 的字节在 *File* 中出现的频数,$i=1,\cdots,10$;$p_i^6=N_i^6\Big/\sum_{i=1}^{194}N_i^6$,$N_i^6$ 为取值为整数集合 S_{194} 中按数值从小到大排序的第 i 个元素的字节在 *File* 中出现的频数,$i=1,\cdots,194$.这里,整数集合 S_{194} 满足

$$S_{194}=\left\{\begin{matrix}0,1,2,\cdots,47,58,59,\cdots,64,91,\\92,\cdots,96,123,124,\cdots,255\end{matrix}\right\}.$$

(16)特征 *Max*+*Ent*

特征 *Max*+*Ent* 维数为 12,是特征 *Max*、*Ent* 的组合,形如$(f_1^M,\cdots,f_6^M,f_1^E,\cdots,f_6^E)$.

(17)特征 *F_512*

512 维特征 *F_512* 形如$(f_1^{512},\cdots,f_{512}^{512})$,计算为

$$f_i^{512}=\frac{N_i^{512}}{l_C},\ i=1,\cdots,512,$$

其中 N_i^{512} 表示在 *File* 中,取值为 $i-1$ 的分块出现的频数.

(18)特征 *F_1024*

1024 维特征 *F_1024* 形如$(f_1^{1024},\cdots,f_{1024}^{1024})$,计算为

$$f_i^{1024}=\frac{N_i^{1024}}{l_C},\ i=1,\cdots,1024,$$

其中 N_i^{1024} 表示在 *File* 中,取值为 $i-1$ 的分块出现的频数.

(19)特征 *F_5Test*

特征 *F_5Test* 维数为 500.在计算中,将 *File* 分为等长的 100 块,并分别开展随机性测试 Frequency、Block Frequency、Runs、Test for the Longest Run of Ones in a block、Cumulative Sums Test[18],得返回值组成的 5 种 100 维特征 *F_F′*、*F_BF′*、*F_R′*、*F_LR′*、*F_CS′*.特征 *F_5Test* 是上述 5 种随机性测试特征的组合,形如

$$(f_1^{F'},\cdots,f_{100}^{F'},f_1^{BF'},\cdots,f_{100}^{BF'},f_1^{R'},\cdots,f_{100}^{R'},$$
$$f_1^{LR'},\cdots,f_{100}^{LR'},f_1^{CS'},\cdots,f_{100}^{CS'}).$$

HUANG Liang-Tao, born in 1993, M. S. candidate. His research interests include machine learning and cryptography.

ZHAO Zhi-Cheng, born in 1992, M. S. candidate. His research interests include applications of probability and statistics in cryptography.

ZHAO Ya-Qun, born in 1961, Ph. D., professor. Her research interests include basic theory in cryptography and applied probability and statistics.

Background

In the field of Cryptanalysis, it's important to recognize the cryptosystem before conducting specific cryptanalysis techniques. Among existing recognition schemes, the research is often restricted in designing single stage schemes or recognizing block ciphers. In this paper, we proposed a definition system for cryptosystem recognition, and suggested a two-stage recognition scheme based on random forest. The scheme consists of 2 sequential stages 'cluster recognition' and 'single recognition' that the cluster of cryptosystem is first recognized and then the exact cryptosystem is identified within the cluster. Extensive works of feature engineering, classifier training and testing have been done to demonstrate the effectiveness of the proposed scheme. Specifically, the recognition accuracy was raised in 3 experimental settings by 19.55%, 21.40%, 22.99%, respectively.

Our work in this paper is a part of a project supported by the open fund from the State Key Laboratory of Information Assurance Technology (No. KJ-15-008), which aims at providing a set of tools for network data analysis. And this paper is a preliminary work in the project so that conducting further analysis on certain kinds of encrypted data is possible. Our team is experienced in Cryptography, Probability theory and Statistics. Our former works including impossible differential cryptanalysis on block ciphers, Boolean functions in cryptography, etc.